(0)

R842-09 ERIC REPORT RESUME

ED 010 211 1-31-67 24 (REV)
SEQUENTIALLY SCALED MATHEMATICS ACHIEVEMENT TESTS--CONSTRUCTION
METHODOLOGY AND EVALUATION PROCEDURES.
GRAHAM, GLENN T.
TGZ65128 UNIVERSITY OF PITTSBURGH, PA.
BR-5-0253-THESIS-3
- -66
EDRS PRICE MF-$0.27 HC-$5.40 135P.

*ACHIEVEMENT TESTS, *TEST CONSTRUCTION, RATING SCALE,
*SEQUENTIAL APPROACH, *TEST VALIDITY, *MATHEMATICS,
ELEMENTARY EDUCATION, GUTTMAN@S SCALOGRAM,
METROPOLITAN ACHIEVEMENT TESTS, PITTSBURGH, PENNSYLVANIA

THIS STUDY APPLIED GUTTMAN@S SCALOGRAM ANALYSIS METHODOLOGY TO TEST CONSTRUCTION, AND DEVELOPED EVALUATION PROCEDURES ON THE RELIABILITY, VALIDITY, AND ITEM ANALYSIS OF THE OBTAINED TESTS. THE METHODOLOGY FOR CONSTRUCTION WAS APPLIED IN FIVE AREAS OF ARITHMETIC ACHIEVEMENT--(1) ADDITION, (2) SUBTRACTION, (3) NUMERATION, (4) TIME-TELLING, AND (5) MONETARY CONCEPTS. SUBJECTS WERE OBTAINED FROM TWO ELEMENTARY SCHOOLS FOR TEST EVALUATIONS. A COMPARISON OF THE SCALED TEST RESULTS TO THE RESULTS OF THE 'METROPOLITAN ACHIEVEMENT TESTS' WAS EMPLOYED. INCONSISTENT RESULTS APPEARED. THE CONCLUSIONS INDICATED THAT IT WAS POSSIBLE TO CONSTRUCT SEQUENTIALLY-SCALED ACHIEVEMENT TESTS IN CERTAIN AREAS OF ARITHMETIC. FURTHER RESEARCH WAS SUGGESTED AT ALL GRADE LEVELS IN SUCH ARITHMETIC ACHIEVEMENT AREAS AS MULTIPLICATION, DIVISION, AND FRACTIONS. (RS)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

5-0253
Thesis No. 3

SEQUENTIALLY SCALED MATHEMATICS ACHIEVEMENT TESTS: CONSTRUCTION METHODOLOGY AND EVALUATION PROCEDURES

RECEIVED
JUL 29 1966
DIVISION OF LABORATORIES
& RESEARCH DEVELOPMENT

By

Glenn Thomas Graham

B.S., University of Pittsburgh, 1962
M.A., University of Pittsburgh, 1965

ED010211

Submitted to the Graduate Faculty of the School
of Education in partial fulfillment of
the requirements for the degree
of Doctor of Education

University of Pittsburgh

1966

FOREWARD

The investigator wishes to express his sincere appreciation to Dr. Richard C. Cox, his thesis advisor, and Dr. C. M. Lindvall, his major advisor, for their assistance and guidance so generously provided both in the preparation of the thesis and throughout the investigator's graduate career. Appreciation is also extended to Dr. John O. Bolvin, Dr. Henry Hausdorff, and Dr. Richard K. Seckinger for their constructive criticism in critically reading this thesis.

The investigator is indebted to the staff of the Learning Research and Development Center, and in particular Dr. Robert Glaser, the director, for the opportunity to employ the Center's resources and facilities during the writing of the thesis.

Gratitude is extended to Mrs. Lois Lackner for her assistance in the construction of the scaled tests.

The investigator is especially grateful to his wife, Bettie, for her faith, patience and encouragement throughout the doctoral program.

The research and development reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education and Welfare, under provisions of the Cooperative Research Program.

TABLE OF CONTENTS

Page

FOREWARD. ii

LIST OF TABLES. iv

LIST OF FIGURES vi

I. INTRODUCTION. 1

II. REVIEW OF RELATED RESEARCH. 6

A. Attempt to Obtain Test Scores Providing Further Information 6
B. Proposed Solution for Interpreting Specific Behaviors from Test Scores 10
C. The Technique of "Scalogram Analysis" 14
D. Criticism of Scalogram Analysis 17
E. Application of Scalogram Analysis 27
F. A Methodology for the Construction of a Sequentially Scaled Achievement Test. . . . 35
G. Evaluation of Sequentially Scaled Achievement Tests 36
H. Summary 39

III. STATEMENT OF THE PROBLEM. 41

IV. PROCEDURE 42

A. Summary 55

V. RESULTS AND DISCUSSION. 58

A. Application of Scaling Methodology. 58
B. Reliability of the Scaled Tests 62
C. Validity of the Scaled Tests. 66
D. Item Analysis of the Scaled Tests 73
E. Comparison of the Scaled Tests and the Metropolitan Achievement Test 78

VI. CONCLUSIONS AND SUGGESTION FOR FURTHER RESEARCH .100

BIBLIOGRAPHY. .108

APPENDIX A. 115

APPENDIX B. 117

APPENDIX C. 121

APPENDIX D. 124

LIST OF TABLES

Table Page

1. REPRODUCIBILITY AND SCALABILITY COEFFICIENTS FOR OAKLEAF SCALES 59

2. REPRODUCIBILITY AND SCALABILITY COEFFICIENTS FOR SICKMAN SCALES 59

3. REPRODUCIBILITY AND SCALABILITY COEFFICIENT FOR THE COMBINED SAMPLE OF PUPILS. 62

4. CORRECTED SPLIT-HALF RELIABILITY COEFFICIENTS FOR THE SCALED TESTS 63

5. SPEARMAN RHO CORRELATION COEFFICIENTS FOR OAKLEAF AND SICKMAN ITEM ORDERS. 65

6. SPEARMAN RHO CORRELATION COEFFICIENTS FOR ITEM ORDERS OF TWO RANDOM SAMPLES FROM THE TESTS BASED ON COMBINED SAMPLES. 65

7. PERCENTAGES OF RAW SCORES EQUALING THE FIRST n ITEMS PASSED (OAKLEAF) 67

8. PERCENTAGES OF RAW SCORES EQUALING THE FIRST n ITEMS PASSED (SICKMAN) 67

9. PERCENTAGES OF RAW SCORES EQUALING THE FIRST n ITEMS PASSED (COMBINED). 68

10. PERCENTAGES OF PREDICTIONS OF PUPIL POSITION IN THE OAKLEAF CURRICULUM SEQUENCE ON THE BASIS OF SCALED TEST RAW SCORE 70

11. NUMBER OF CURRICULUM SKILLS PER LEVEL FOR TESTED UNITS 72

12. RATIOS OF NUMBER OF TEST ITEMS TO NUMBER OF CURRICULUM SKILLS IN LEVELS C, D, and E. 72

13. ITEM ANALYSIS FOR OAKLEAF SCALED TESTS 74

14. ITEM ANALYSIS FOR SICKMAN SCALED TESTS 75

15. ITEM ANALYSIS FOR COMBINED SCALED TESTS. 76

16. ORIGINAL ITEM ORDERS FOR POOR ITEMS. 77

17. SPLIT-HALF RELIABILITY COEFFICIENTS FOR SCALED TESTS AND METROPOLITAN ACHIEVEMENT TESTS 79

Table Page

18. UPPER 27 PERCENT--LOWER 27 PERCENT ITEM ANALYSIS FOR OAKLEAF SCALED TESTS. 81

19. ITEM DESIGNATED AS POOR BY TWO ITEM ANALYSIS PROCEDURES: THE SCALE PROCEDURE AND UPPER-LOWER 27 PERCENT 83

20. REPRODUCIBILITY AND SCALABILITY OF TESTS DERIVED FROM THE METROPOLITAN ACHIEVEMENT TESTS. 86

21. ANALYSIS OF NUMBER OF RAW SCORE PATTERNS PER TEST 88

22. ANALYSIS OF SCORE PATTERNS PER TEST FOR SCALES DERIVED FROM THE METROPOLITAN TESTS. 90

23. PERCENTAGES OF PREDICTIONS OF PUPIL POSITION IN THE OAKLEAF CURRICULUM SEQUENCE: SCALED TEST RAW SCORE AND METROPOLITAN TEST RAW SCORE. 92

24. NUMBER OF ITEMS TESTING EACH LEVEL OF THE FIVE SELECTED UNITS IN THE OAKLEAF CURRICULUM: SCALED TESTS VERSUS METROPOLITAN TESTS 96

LIST OF FIGURES

Figure **Page**

1. **PARALLELOGRAM PATTERN OF A PERFECT FIVE ITEM SCALE 15**

I. INTRODUCTION

The score obtained by an individual on an achievement test will, depending on the standard employed, generally provide two types of information. If information pertaining to an individual's standing in reference to others in a particular group is desired, a relative standard is employed. Glaser referred to such a measure as a "norm-referenced measure."[1] Scores on norm-referenced measures are typically in the form of percentiles, equivalent scores, standard scores, etc.

If information pertaining to an individual's level of mastery of some specified criterion is desired, an absolute standard is employed. Such a measure is referred to by Glaser as a "criterion-referenced measure."[2] The same distinction has been made by Ebel with "Normative Standard Scores" versus "Content Standard Scores,"[3] and Flanagan with "standard" versus "norm" scores.[4] According to Glaser:

---

[1] Robert Glaser, "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, XVIII (1963), 520.

[2] Ibid., p. 519.

[3] Robert L. Ebel, "Content Standard Test Scores," Educational and Psychological Measurement, XXII (1962), 15.

> Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others.[5]

In many classroom situations the use of norm-referenced measures has been emphasized. Typically, the instructional sequence, materials, and rate of progress for each student in the class are held constant. At the end of some specified unit an achievement test is administered to the entire class at the same time. The student's scores are then ranked in relation to each other, or in some instances, the scores are ranked and interpreted in reference to some normative group.

While providing information concerning the number of right and wrong answers and the relative standings of individuals, the score on a norm-referenced test does not indicate what specific behaviors the student has mastered. Except in the extreme cases where every item is passed or failed, raw scores or percentages indicate only the number of questions answered correctly. Converting the raw scores to percentiles, standard scores, equivalent scores, etc., still provides no information concerning the particular skills the student has or has not mastered. "From a percentile we know the location of a pupil's score in the distribution of scores of the normative pupils, we still do not know how much arithmetic a pupil understands."[6] Different scores indicate that different items have been answered correctly, but not what items were answered correctly. The same scores do not necessarily indicate that the same items have been passed; success on many different items has probably occurred. To determine the specific behaviors which have been mastered, the individual items need to be examined.

---

[4] John C. Flanagan, "Units, Scores, and Norms," Educational Measurement, E. F. Lindquist, editor (Washington, D.C.: American Council on Education, 1951), p. 698.

[5] Glaser, op. cit., p. 520.

In many cases knowing the ranks of the individuals is sufficient. But with the development of programs of individualized instruction, such as programed learning or nongraded classrooms, criterion-referenced measures have become increasingly important. In individualized instruction each student sets his own pace for learning and in the process may pursue varied curriculum sequences and materials. The performance criteria for a specified unit of work may be identical for all students however, their performance being compared to an absolute rather than a relative standard. Minimum levels of mastery are established which the student must meet before progressing to the next unit. Tests for units are not administered to the group as a whole, but to individuals as they complete these units in their instructional sequence. The score on this type of test is used to determine whether a student progresses to the next unit. His score is compared to the criterion established for the unit, not to the scores of others in the group.

---

[6] Fred T. Tyler and Walter R. Stellwagen, "The Search for Evidence about Individual Differences," Individualizing Instruction, Sixty-first Yearbook of the National Society for the Study of Education, Part I (Chicago, Illinois: The University of Chicago Press, 1962), p. 99.

The "impact of individualized instruction on testing and diagnosis in the schools" has been discussed by Coulson and Cogswell.[7] They stated that the trend toward individualized instruction "... is not an isolated phenomenon, independent of other educational activities such as testing...."[8] The authors spoke of the need to develop "... a go/no go test determining whether a student is ready to graduate or to progress to the next study unit...."[9] If such diagnostic techniques as Coulson and Cogswell described could be developed, the authors suggested that "they should provide not only a means for more effective instruction, but also a basis for constructing more useful theories of education and learning."[10]

---

[7] John E. Coulson and John F. Cogswell, "Effect of Individualized Instruction on Testing," Journal of Educational Measurement, II (1965), pp. 59-64.

[8] Ibid., p. 59.

[9] Ibid.

[10] Ibid., p. 63.

The need for further consideration in test development has also been recommended by Glaser who stated:

> Test development has been dominated by the particular requirements of predictive, correlational aptitude test 'theory.' Achievement and criterion measurement has attempted frequently to cast itself in this framework. However, many of us are beginning to recognize that the problems of assessing existing levels of competence and achievement and the conditions that produce them require some additional consideration.[11]

Since criterion-referenced measures are directly concerned with "assessing existing levels of competence and achievement," they should provide information concerning the students' successes and failures on specified behaviors. Whether this information can be obtained from the raw score on a criterion-referenced test, or whether, as in the case of norm-referenced tests, the individual test items require examination remains a problem.

---

[11] Glaser, *op. cit.*, p. 521.

II. REVIEW OF RELATED RESEARCH

A. Attempts to Obtain Test Scores Providing Further Information

Attempts have been made to provide scores yielding further information than that furnished by norm-referenced scores. Grossnickle employed Thurstone's paired comparisons technique to investigate the possible scaling of individuals making certain test scores.[12] The desire was to obtain scores which would remain relatively the same for individuals regardless of the group in which they were placed. A biology test of 100 items was administered to 100 persons whose scores were then ranked from highest to lowest. These scores were then grouped by 5's to form twenty "hypothetical individuals," the top five scores being individual number 1, etc. Scaled scores for these twenty individuals were obtained.

A new group of thirty persons was selected and given the test. These persons were also combined into six "individuals," and scaled scores obtained. Four "individuals" from the original group were then randomly selected and combined with the latter group of six. While the four scaled scores did not remain the same in the new group as in the old, the distance between the scores remained stable. Grossnickle concluded that "this experiment using the paired comparison method, has proved that it is possible to scale individuals taking any mental and educational test."[13]

---

[12] Louise T. Grossnickle, "The Scaling of Test Scores by the Method of Paired Comparisons," Psychometricka, VII (1942), pp. 43-64.

[13] *Ibid*., p. 62.

Such a conclusion seems somewhat unwarranted. The author claimed this truth for individuals, yet she never dealt with individuals; also, she generalized to the population of *all* mental and educational tests from one biology test. No additional meaning could really be attached to the obtained scores since they changed depending on the reference group.

A further attempt to add meaning to test scores was reported by Tucker at the 1952 Invitational Conference on Testing Problems. According to Tucker:

> ... experimental and analytic methods for test development and score scaling may exist or be developed which do not depend on the relative number of examinees who receive each particular score in a reference group of examinees.[14]

In keeping with this suggestion Tucker attempted to obtain scores relating individuals' proficiency on a skill to the difficulty of a task performed at a marginal degree of success.[15] He provided the following example: In receiving telegraph code an individual will make fewer errors receiving slow signals than fast signals. At some speed he would receive with 90% accuracy. This signal speed could be used to characterize that individual's level of proficiency. Tucker proposed a system for defining a scale of difficulty for intellectual skills.

---

[14] Ledyard R. Tucker, "Selecting Appropriate Score Scales for Tests—Scales Minimizing the Importance of Reference Groups," *Proceedings, 1952 Invitational Conference on Testing Problems* (Princeton: Educational Testing Service, 1953), pp. 27-28.

[15] Ledyard R. Tucker, "A Level of Proficiency Scale for a Unidimensional Skill," *American Psychologist*, VII (1952), 408. (Abstract)

Such a model involved several steps which included establishing subgroups of individuals with approximately equal skills, obtaining the proportion of successes on each task for each subgroup, and determining a scale value for each subgroup. Tucker reported that results from an application to a set of verbal analogy items indicated promising possibilities, but he provided no data. He proposed a further tryout involving the scaling of vocabulary items from fourth grade through college. It would appear, however, that this technique will provide a score similar to a mental age, rather than indicating what specific behaviors have been mastered.

A similar attempt to provide scores indicative of a level of proficiency has been reported by Ebel. He discussed two studies concerned with providing "content standard scores." These scores are based directly on the tasks which compose the content of the test, and are defined as a "percent of a systematic sample from a defined domain of tasks which an individual has performed successfully."[16]

Ebel constructed a test of knowledge of word meanings based on a sample of 100 words from a specified dictionary. The words were arranged in alphabetical order; the task was to match the words with their corresponding definitions. Ebel stated that "these tests constitute one operational definition of the proportion of words in a certain dictionary for which a person 'knows' the meaning...."[17]

---

[16] Ebel, "Content Standard Test Scores," Educational and Psychological Measurement, XXII (1962), 15.

[17] Ibid., pp. 24-25.

Ten items were also selected by Ebel from the mathematics section of the 1959 Preliminary Scholastic Aptitude Test. Initially all fifty items of the test were classified into ten content categories, e.g., "Calculations with fractions, Verbal problems, Percentage and statistics," etc. The discriminating power of each item was found by subtracting the proportion of correct responses in a low scoring group (PSAT scores below 300) from the proportion of responses in a high scoring group (PSAT scores above 700). The item in each category with the highest discriminating power was chosen. These items were then scored on six sets of 100 answer sheets which had PSAT scores near 750, 650, 550, 450, 350, and 250. The most frequent score on the ten items was found for each group. For example, a score of 9 was the most frequent for those scoring 750; a score of 3 was most frequent for those scoring 450. Therefore, a score on the ten item test was taken as an indication of the score on the PSAT.

In each of these two examples the test provided information related to content, however, no information relating to the mastery of specific skills was obtained. In the former the score indicated a percentage of the words, in the latter the score indicated the most frequent score obtained for a given group, but did not indicate what score a given individual would obtain nor to what items that score pertained. Also, in reference to the latter of Ebel's examples, the items were chosen to discriminate between scores of 700 and 300. The author evidently assumed that these same items would have the best discriminating powers for the other groups reported.

B. Proposed Solutions for Interpreting Specific Behaviors from Test Scores

One solution to obtaining a criterion-referenced test whose score would indicate specific behaviors mastered by a student would be a test whose items were sequentially scaled. The items would be so arranged that once an individual missed an item he would miss all subsequent items in the test. For example, an individual obtaining a score of 8 would have answered each of the first eight items correctly and none of the items beyond 8 correctly; a student failing item 3 should fail all subsequent items. According to Ebel:

> It is possible to imagine a test which would give highly consistent results across items and across students when administered to a particular group. Results would be called consistent if success by a particular student on a particular item practically guaranteed success on all other items in the test which were easier for the group than that item. Correspondingly, failure on a particular item would almost guarantee failure on all harder items if the student responses were highly consistent.... Such tests can be imagined but are seldom met with in practice.[18]

Two techniques relating to scaled tests are Loevinger's "homogeneous tests"[19] and Guttman's "scalogram analysis."[20]

---

[18] Robert L. Ebel, Measuring Educational Achievement (Englewood Cliffs: Prentice-Hall, Inc., 1965), pp. 361-62.

[19] Jane Loevinger, "A Systematic Approach to the Construction and Evaluation of Tests of Ability," Psychological Monographs, LXI (1947), No. 285; Jane Loevinger, "The Technic of Homogeneous Tests Compared with Some Aspects of 'Scale Analysis' and Factor Analysis," Psychological Bulletin, XLV (1948), pp. 507-29.

[20] Louis Guttman, "A Basis for Scaling Qualitative Data," American Sociological Review, IX (1944), pp. 139-150;

Loevinger defined "perfectly homogeneous tests" of ability as tests "such that, if A's score is greater than B's score, then A has more of some ability than B, and it is the same ability for all individuals A and B who may be selected."[21] She proposed a "coefficient of homogeneity," ranging in value from zero to one, which is the ratio of the difference between the variance of a given test and the variance of a "perfectly heterogeneous test" with the same distribution of item difficulties, to the difference between the variance of a perfectly homogeneous test with the same distribution of item difficulties and the perfectly hetergeneous test:

$$\text{Homogeneity } (H_1) = \frac{Vx - Vhet}{Vhom - Vhet}$$

A "perfectly heterogeneous" test is defined as a test "composed of items each of which measures an ability independent of the abilities measured by the other items."[22]

What Loevinger desired was a test which was consistent with respect to the ability being measured. People who obtained the same scores would have answered the same items.

Louis Guttman, "The Problem of Attitude and Opinion Measurement," in Samuel A. Stouffer *et al.*, *Measurement and Prediction* (Vol IV of *Studies in Social Psychology in World War II*. 4 vols.; Princeton: Princeton University Press, 1950), pp. 46-59.

[21] Loevinger, "A Systematic Approach," p. 28.

[22] *Ibid.*

correctly. While being at different levels of difficulty, the items of the test had to measure the same content that defined the ability. Behavior could be inferred from test score by applying one of Loevinger's theorems, "When the items of a perfectly homogeneous test are arranged in order of increasing difficulty, every individual will pass all items up to a certain point and fail all subsequent items."[23]

Of concern to Loevinger was whether the proposed estimate of homogeneity was unbiased. Some evidence that the estimate may be biased was provided by Carroll.[24] Employing random numbers and hypothetical individuals he found Loevinger's coefficient "to be biased positively because of chance variations in item difficulties."[25] To be homogeneous in the Loevinger model the items should measure the same ability but at varying levels of difficulty. Carroll, however, was able to obtain a value as high as .32 for Loevinger's coefficient of homogeneity with items varying in difficulty only by chance.

The Loevinger technique is limited, however, to tests composed of items of the same content. In most tests of ability the content is likely to vary within the test.

---

[23] Ibid.; Loevinger, "The Technic of Homogeneous Tests," p. 508.

[24] John B. Carroll, "Criteria for the Evaluation of Achievement Tests—from the Point of View of Their Internal Statistics," Proceedings, 1950 Invitational Conference on Testing Problems (Princeton: Educational Testing Service, 1951), pp. 95-99.

[25] Ibid., p. 97.

The second technique, Guttman's "scalogram analysis" does not require items of the same content. According to Suchman:

> It is also important to remember that scale analysis should not be depended upon to determine content. An item of differing content may fit into the scale pattern of an area, while items with homogeneous content need not scale.[26]

Edwards and Kilpatrick have also noted this characteristic of scalogram analysis:

> The merits of scale analysis, as a _technique for evaluating_ a set of items, are obvious and need no defense. But scale analysis can be applied to any set of items, regardless of how the set is selected.[27]

Arising from problems in attitude scaling and opinion polling, scalogram analysis attacks directly the problem of determining behavior from score. As Guttman stated:

> From a person's score we would then know precisely to which problems he knows the answers and to which he does not know the answer. Thus a score of 2 does not mean simply that the person got two questions right, but that he got two particular questions right, namely, the first and second. A person's behavior on the problems is reproducible from his score.[28]

---

[26] Edward A. Suchman, "The Scalogram Board Technique for Scale Analysis," in Samuel A. Stouffer _et al._, _Measurement and Prediction_ (Vol. IV of _Studies in Social Psychology in World War II_. 4 vols,; Princeton: Princeton University Press, 1950), p. 119.

[27] Allen L. Edwards and Franklin P. Kilpatrick, "Scale Analysis and the Measurement of Social Attitudes," _Psychometrika_, XIII (1948), 109.

[28] Guttman, "A Basis for Scaling," p. 143.

Guttman further discussed the possibility of determining behavior from score when he stated:

> Scale analysis tests the hypothesis that a group of people can be arranged in an internally meaningful rank order with respect to an area of qualitative data. A rank order of people is meaningful if, from the person's rank order, one knows precisely his responses to each of the questions or acts included in the scale.[29]

C. The Technique of "Scalogram Analysis"

Guttman defines a scale as each person's responses being reproducible from his rank alone.[30] The technique for determining the existence of a scale involves essentially two steps: (1) ranking the items from "most extreme" to "least extreme," i.e., the item chosen or answered correctly by the fewest people ("the most extreme") to the item chosen or answered correctly by the most people ("the least extreme"); and (2) ranking the individuals from lowest to highest on the basis of total score. If a scale exists the resultant pattern when correct and incorrect responses are tabulated will be a parallelogram (or a triangle if only correct responses are recorded).[31] The following example will provide an illustration of the resultant patterns. Consider a five item test administered to five people and found to be a perfect scale. Figure 1 shows the parallelogram pattern when both correct and incorrect responses are recorded (X's).

---

[29] Louis Guttman, "The Basis for Scalogram Analysis," in Stouffer *et al.*, *Measurement and Prediction*, p. 88.

[30] *Ibid.*, p. 62.

[31] *Ibid.*, pp. 60-90.

ITEMS

	Incorrect					Correct				
Individuals	1	2	3	4	5	1	2	3	4	5
1						X	X	X	X	X
2					X	X	X	X	X	
3				X	X	X	X	X		
4			X	X	X	X	X			
5		X	X	X	X	X				

FIGURE 1

PARALLELOGRAM PATTERN OF A PERFECT FIVE ITEM SCALE

In discussing the rank ordering of individuals and items from such a pattern Suchman said:

> Such a rank order has the property of permitting one to derive from the rank order the exact characteristics of the individuals in that rank since there is only one possible combination of items for any single rank order. Furthermore, the rank order has the quality that any individuals in a higher rank possess all the characteristics of the individuals in a lower rank, and at least one more in addition.[32]

The above description pertains to a perfect scale, that is, each individual's response to each item can be perfectly reproduced. Such perfect scales are usually not found in practice, just as perfectly reliable tests are not found in practice. The degree to which the instrument approximates a perfect scale is measured by a "coefficient of reproducibility." The coefficient is defined as the "empirical relative frequency with which values of the attributes do correspond to intervals of a scale variable."[33] Thus, the coefficient provides an indication of how well an individual's response pattern can be reproduced knowing his total score. The value of .90 was arbitrarily established as an acceptable lower limit of the coefficient.

---

[32] Edward A. Suchman, "The Logic of Scale Construction," Educational and Psychological Measurement, X (1950), 90.

As described by Guttman in his original article[34] and again by Suchman in 1950,[35] scalogram analysis was performed by the use of scalogram boards. These are devices which, through the use of balls and slats, permit the shifting of the rank orders of individuals and items (or the combination of categories for items with multiple categories) to obtain the best scale.

The use of scalogram boards has not always been feasible however; the cost of the boards is prohibitive, they have a fixed capacity, and have been termed cumbersome.[36] In answer to such criticism Guttman devised a paper and pencil technique, the "Cornell Technique" to supplant the scalogram boards.[37] The technique is applied to the data in the same way as the boards, shifting the rank orders to obtain the best arrangement of items. In either case, whether applying scalogram boards or the Cornell Technique, the reproducibility coefficient is computed in the same manner: (1) errors are tabulated by counting the number of responses occurring outside the cut-off points for each score; (2) the errors are divided by the total number of possible responses (number of people x number of items), and (3) the obtained quotient is subtracted from 1.[38]

---

[33] Guttman, "The Basis for Scalogram Analysis," p. 89.

[34] Guttman, "A Basis for Scaling," pp. 139-150.

[35] Suchman, "The Scalogram Board Technique," in Stouffer et al., Measurement and Prediction, pp. 91-121.

[36] Wilfred A. Gibson, "A Simple Procedure for Rearranging Matrices," Psychometrika, XVIII (1953), pp. 111-113; Leon Festinger, "The Treatment of Qualitative Data by 'Scale Analysis,' Psychological Bulletin, XLIV (1947), pp. 149-161.

$$\text{Reproducibility (R)} = 1 - \frac{\text{Errors}}{\text{No. Items x N}}$$

D. Criticisms of Scalogram Analysis

The use of the reproducibility coefficient has been criticized in the literature. Several authors have found the coefficient to be arbitrary and to be affected by the difficulty levels of the test items. The major objection is that the reproducibility coefficient can be spuriously high because of items with extreme marginal frequencies.[39] Guttman was aware of the effect of extreme marginals on the reprocibility coefficient, however, and never stated that a high reproducibility coefficient was the criterion for scalability. "Reproducibility by itself is not a sufficient test of scalability. It is the principal test, but there are at least four other features that should be taken into account...."[40]

---

[37] Louis Guttman, "The Cornell Technique for Scale and Intensity Analysis," *Educational and Psychological Measurement*, VII (1947), pp. 247-280.

[38] Guttman, "The Basis for Scalogram Analysis," in Stouffer *et al.*, *Measurement and Prediction*, p. 77.

The four additional criteria which are employed to insure against spurious reproducibility are: (1) The number of items in the test should exceed 10. (2) The more categories that could remain uncombined, the more credible the inference of scalability; this criterion does not apply to dichotomous items. (3) The marginal distributions should contain as wide a range as possible, but with few extreme marginals. In the case of dichotomous items an extreme marginal would be a category chosen by 80% or more of the individuals. The reproducibility of an item can never be less than the frequency of the most frequently chosen category. (4) The errors should not fall into a pattern, i.e., there should not be a series of persons, all having identical errors.[41]

---

[39] Festinger, "The Treatment of Qualitative Data," pp. 149-161; H. J. Eysenck and S. Crown, "An Experimental Study in Opinion-Attitude Methodology," International Journal of Opinion and Attitude Research, III (1949), pp.47-86; H. J. Eysenck, "Measurement and Prediction; A Discussion of Volume IV of Studies in Social Psychology in World War II: I.," International Journal of Opinion and Attitude Research, V (1951), pp. 95-102; Benjamin W. White and Eli Saltz, "Measurement of Reproducibility," Psychological Bulletin, LIV (1957), pp. 81-99.

[40] Guttman, "The Basis for Scalogram Analysis," p. 78.

A single criterion to evaluate the spuriousness of the reproducibility coefficient has been suggested by Menzel.[42] He developed a "coefficient of scalability" having the following advantages: (1) it provides a safeguard against spuriousness without relying on extraneous rules, (2) it does not introduce the judgment of the investigator in applying a rule, and (3) it permits analysis of scalograms that had to be rejected because of extreme marginals. The coefficient may, therefore, show that high scalability exists in spite of extreme marginals.

The coefficient is computed by (1) obtaining the errors as in the computation of the reproducibility coefficient, (2) for dichotomous items—summing the non-modal marginal frequencies for items and for individuals, and taking the smaller of the two scores, and (3) dividing the errors by the sum obtained in step 2, and subtracting the resultant quotient from 1. A minimum value of .65 was established as the criterion for scalability. Among those who have recommended and employed the coefficient of scalability in conjunction with the reproducibility coefficient are Auld, Eron, and Laffal; Lesser; and Pearson.[43]

---

[41] Guttman, "The Basis for Scalogram Analysis," pp. 78-80.

[42] Herbert Menzel, "A New Coefficient for Scalogram Analysis," _Public Opinion Quarterly_, XVII (1953), pp. 268-280.

Not meeting the criteria for scalability in scalogram analysis should not be interpreted as showing the non-existence of a scale according to Eysenck and Crown. "Ultimately we shall achieve the position of the physicist whose scales show approximately 100% reproducibility, [but] there is little to be gained by decrying the very real usefulness of many of our own present-day scales."[44] In essence these authors are arguing _for_ use of less reproducible scales, but not arguing against highly reproducible scales.

Continuing the criticisms of the reproducibility coefficient, the reason for using the reproducibility coefficient was questioned by Davis when he stated, "We compute a reproducibility coefficient not because we have any real desire to reproduce response patterns from scale scores but, rather, because we hope that it is an index of certain measurement properties."[45] Contrary to Davis' assumptions, however, reproducing responses from scores is the goal established in the present study.

---

[43] Frank Auld, Jr., L. A. Eron, and J. Laffal, "Application of Guttman's Scaling Method to the T. A. T.," _Educational and Psychological Measurement_, XV (1955), pp. 422-435; Gerald S. Lesser, "Application of Guttman's Scaling Method to Aggressive Fantasy in Children," _Educational and Psychological Measurement_, XVII (1958), pp. 543-551; Richard G. Pearson, "Plus Percentage Ratio and the Coefficient of Scalability," _Public Opinion Quarterly_, XXI (1957), pp. 379-380.

[44] Eysenck and Crown, _op_. _cit_., p. 66.

Smith disclaims the notion of reproducibility in testing for the existence of any scale. He reached this conclusion because he obtained two group factors on factor-analyzing the data reported by Guttman in the 1947 article concerning the Cornell Technique. According to Guttman only one factor should have been found.[46] Guttman's reply was that Smith's "numerical work cannot be anything but pure nonsense."[47] Guttman showed that Smith reported perfect correlations between items 2 and 4, and between items 3 and 5, yet 2 correlated differently with the other items than 4, and 3 correlated differently with the other items than 5. Smith's matrix was non-Gramian and as a result could not be factored by the Thurstone Technique which he employed.

In addition to the coefficient of reproducibility, other aspects of scalogram analysis have been subjected to criticism. The determination of cut-off points for scores has been found by some authors to be arbitrary and difficult.[48] Their criticisms have pertained to attitude scales having items with three or more response categories. In an achievement test having dichotomous items, where the desire is to infer behavior from score, the various scores would determine the cut-off points. Theoretically an individual should answer items to a certain point and then stop. Therefore, a score of 4 would cut off the first four items, etc. As a result, the above criticisms would not apply.

---

[45] James A. Davis, "On Criteria for Scale Relationships," American Journal of Sociology, LXII (1958), 374.

[46] R. G. Smith, Jr.,"'Randomness of Error' in Reproducible Scales," Educational and Psychological Measurement, XI (1951), pp. 587-596.

[47] Louis Guttman, "On Smith's Paper on 'Randomness of Error' in Reproducible Scales," Educational and Psychological Measurement, XIII (1953), 507.

Contrary to the above example, opinions have been expressed that individuals do not usually answer items up to a certain point and then stop. Rather, a gradual transition from correct to incorrect has been suggested. Brown, Bartelme, and Cox proposed that "the score of the individual is then at that point on the scale at which the average deviation of the right items above it equals the average deviation of the wrong items below it."[49] The authors based their conclusions on results obtained from the Gesell Development Schedule. Glaser hypothesized that on certain tests measuring one dimension, when the test items are ordered in terms of their scale position, there is a region of transition from positive to negative responses.[50] He further hypothesized that the distribution of inconsistent responses in this region is approximately normal. Glaser analyzed the following tests: The Faust-Schorling Test Test of Functional Thinking in Mathematics, the Differential Aptitude Space Relations Test, and a vocabulary test composed of items from the Stanford-Binet, Wechsler-Bellevue, Wide Range, and Columbia Vocabulary tests. Each test was composed of 80 items. The results showed the distribution for the vocabulary test to be approximately normal. The distributions for the mathematics and space relation tests were truncated. Glaser attributed the truncated distribution to the restricted range of test items.[51] If more items at higher levels of difficulty could have been added the distribution of responses might well have approximated normality.

---

[48] K. E. Clark and P. H. Kriedt, "An Application of Guttman's New Scaling Techniques to an Attitude Questionnaire," Educational and Psychological Measurement, VIII (1948), pp. 215-223; Allen L. Edwards, "On Guttman's Scale Analysis," Educational and Psychological Measurement, VIII (1948), pp. 313-318; Edgar F. Borgatta and David G. Hays, "Some Limitations on the Arbitrary Classification of Non-Scale Response Patterns in a Guttman Scale," Public Opinion Quarterly, XVI (1952), pp. 410-416.

[49] C. W. Brown, P. Bartelme, and G. M. Cox, "The Scoring of Individual Performance on Tests Scaled According to the Theory of Absolute Scaling," Journal of Educational Psychology, XXIV (1933), 655.

The results should be interpreted in the light of the type of test employed in each of the above studies. All tests were published tests which were not constructed to yield scores from which behavior could be inferred. In addition, the range of gradual transition from pass to fail could well be attributed to very gradual transitions in item difficulties accompanied by a large number of items at each difficulty level. It could be hypothesized that as the number of items increases, and the steps between item difficulties becomes more gradual, the distribution of inconsistent responses would approach normality. The studies cited above lend some support for this. The opposite could also be hypothesized, i.e., with a decreasing number of items and with more discrete steps between difficulties, the distribution of inconsistencies would depart from normality. The truncated distributions obtained by Glaser offer some evidence in support of this. Carrying the latter hypothesis to its extreme, it could be hypothesized that at some point individuals would no longer have inconsistent responses but _would_ answer items to a certain point and then stop. It is this type of test that is suggested in the present study, and scalogram analysis is suggested as a technique to yield such a test.

---

[50] Robert Glaser, "Multiple Operation Measurement," Psychological Review, LVII (1950) pp. 241-253; Robert Glaser, "The Application of the Concepts of Multiple-Operation Measurement to the Response Patterns on Psychological Tests," Educational and Psychological Measurement, XI (1951), pp. 372-382.

[51] Glaser, "The Application of the Concepts," p. 375.

Scalogram analysis has also been criticized by some as inadequate for the selection of items.[52] The reply to these findings is simply that scalogram analysis is not an item selection technique. As Guttman states, "We have continually stressed that items are to be selected before any statistical analysis is performed, and are not to be rejected because of any statistical analysis.... Scale analysis is not a technique for item selection and rejection, but rather for studying the structure of a universe...."[53]

---

[52] Edwards, "On Guttman's Scale Analysis," pp. 313-318; Edwards and Kilpatrick, "Scale Analysis," pp. 99-114; Allen L. Edwards and Franklin P. Kilpatrick, "A Technique for the Construction of Attitude Scales," _Journal of Applied Psychology_, XXXII (1948), pp. 374-384; P. H. Kriedt and K. E. Clark "'Item Analysis' Versus 'Scale Analysis,'" _Journal of Applied Psychology_, XXXII (1949), pp. 114-121.

For Guttman, the universe "is the concept whose scalability is being investigated, such as marital adjustment, opinion of British fighting ability, knowledge of arithmetic, etc. The universe consists of all the attributes that define the concept."[54] One aspect of the theory of scalogram analysis is that from a sample of items comprising the scale "inferences can be drawn concerning the complete distribution of the universe for the population.... The hypothesis that the complete distribution is scalable can be adequately tested with a sample distribution."[55] Criticisms of this aspect seem, to this investigator, to be warranted. It would appear that while a sample of items could well be scaled, as for example, in the following 3 item test:

---

[53] Louis Guttman, "Measurement and Prediction; a Discussion of Vol. IV of Studies in Social Psychology in World War II: II. Scale Analysis, Factor Analysis, and Dr. Eysenck," International Journal of Opinion and Attitude Research, V (1951), 112.

[54] Guttman, "The Basis for Scalogram Analysis," in Stouffer et al., Measurement and Prediction, p. 80.

[55] Ibid., p. 89.

$$2 + 2 = \underline{\hspace{2cm}} \quad \sqrt{2} = \underline{\hspace{2cm}} \quad \int 2x\,dx = \underline{\hspace{2cm}} ,$$

the conclusion that the universe of mathematics is scalable is not tenable. Many skills in mathematics depend on the order taught; many skills are parallel, being of the same difficulty. Schuessler argued that the sample results are both a function of the way the universe is defined by the investigator and the manner in which the items are chosen from a field of content defining the topic.[56] This implies that conclusions concerning the scalability of a universe may be restricted to a given investigator's version.

Further criticism of this aspect of scalogram analysis was provided by Torgerson[57] and Campbell and Kerckhoff.[58] Each warned against generalizing to the universe from a sample. Campbell and Kerckhoff stated that the proposition, "If a universe is scalable any sample selected from the universe will be scalable," is not identical to the proposition, "If a sample is scalable the universe from which the sample is selected is scalable." The authors suggested that if the latter proposition is warranted any other items from the same universe should also scale with the original set. They reported, however, that judges have not been consistent in making these additional selections, but, unfortunately, no empirical evidence was provided.

---

[56] Karl Schuessler, "Item Selection in Scale Analysis," *American Sociological Review*, XVII (1952), pp. 183-192.

[57] Warren S. Torgerson, *Theory and Methods of Scaling* (New York: John Wiley & Sons, Inc., 1958).

[58] Ernest Q. Campbell and Alan C. Kerckhoff, "A Critique of the Concept 'Universe of Attributes,'" *Public Opinion Quarterly*, XXI (1957), pp. 295-303.

The above criticisms are concerned, however, with relating a sample to the universe, not with scaling a sample of items by applying the scalogram technique. A test is considered as being composed of a sample of items representing the population of possible items pertaining to a given area. The proposed study is to determine if a test will scale, if the test will yield a score from which behavior can be inferred. If the Guttman scalogram technique can be applied to achievement testing in order to obtain such scores, the ensuing conclusions will be concerned with the sample of items, the test, not to the universe represented by the sample of items.

E. Applications of Scalogram Analysis

The application of scalogram analysis to achievement testing has been suggest by Guttman on several occasions. "Scale analysis is applicable much more widely than to attitudes. For example, it is useful for mental tests and examinations."[59]

---

[59] Louis Guttman, "The Principal Components of Scalable Attitudes," Mathematical Thinking in the Social Sciences, Paul F. Lazarsfeld, editor (Glencoe, Illinois: Free Press, 1954), p. 217

He also described its use "with large classes of behavior like achievement tests."[60] "For achievement tests, where all items are dichotomous—being marked either right or wrong—the Cornell technique is perhaps the best of all to be used."[61]

While Guttman has suggested that scalogram analysis be applied to achievement testing it has been employed almost entirely in other areas. The most widespread application has been in the areas of opinion and attitude measurement. In Volume IV, Studies in Social Psychology in World War II, Guttman refers to at least seven different studies related to various attitudes of soldiers during the Second World War. Among the many attitude studies reported in the literature, a brief list includes: (1) Niven's comparison of the Cornell Technique and the Reciprocal Averages technique to the attitudes of manufacturing supervisors,[62] (2) an investigation of attitude toward economic liberalism—conservatism,[63] (3) opinion toward science,[64] (4) attitudes toward negroes,[65] (5) the development of an attitude scale on anti-semitism,[66] (6) the scaling of interview responses bearing on the favorability of attitude toward marriage,[67] and (7) Dodd's application to opinion polls in general.[68]

---

[60] Guttman, "The Basis for Scalogram Analysis," in Stouffer et al., Measurement and Prediction, p. 61.

[61] Louis Guttman, "On Festinger's Evaluation of Scale Analysis," Psychological Bulletin, XLIV (1947), 458.

[62] Jarold R. Niven, "A Comparison of Two Attitude Scaling Techniques," Educational and Psychological Measurement, XIII (1953), pp. 65-76.

[63] Clark and Kriedt, op. cit., pp. 215-223.

[64] Edwards and Kilpatrick, op. cit., pp. 374-384.

[65] Kriedt and Clark, op. cit., pp. 114-121.

Other areas have also utilized application of the Guttman technique. Its successful application to projective techniques has been shown by Auld, Eron, and Laffal[69] and Lesser.[70] Auld, *et al*., applied scalogram analysis to themes from four selected pictures of the *Thematic Apperception Test* given to 100 sailors attending submarine school. While the authors did not succeed in constructing a scale of aggression, they did succeed in constructing a scale of sexual motivation. Lesser applied the scaling procedure to the fantasy aggression responses of a sample of pre-adolescent boys. Again the criteria for scalability were met.

---

[66] Eysenck and Crown, "An Experimental Study," pp. 47-86.

[67] Robert McGinnis, "Scaling Interview Data," *American Sociological Review*, XVIII (1953), pp. 514-521.

[68] S. C. Dodd, "A Simple Test for Predicting Opinions from Their Subclasses," *International Journal of Opinion and Attitude Research*, II (1948), pp. 1-25.

[69] Auld, Eron, and Laffal, "Application of Guttman's Scaling to the T.A.T.," pp. 422-435.

[70] Lesser, "Application of Guttman's Scaling Method to Aggressive Fantasy," pp. 543-551.

In addition to projective techniques, scalogram analysis has had application in other diversified areas. Riley et al., used the techniques to scale groups and objects of group action.[71] For example, in a group certain pairs talk of movies, others of movies and petting, but no groups talk of petting alone. Movies and petting were scaled on the degree of intimacy which they represented as subjects for conversation. The application of scale analysis to the scaling of objects was also reported by Abell.[72] Through the use of questionnaire items dealing with homemaking practice, Abell found that foods served, use of preservatives, and vegetables grown were scaled.

Kofsky found tasks involving the classification of objects to be scaled for children of ages four to nine.[73] The classification schemes were based on the developmental sequence of cognitive skills hypothesized by Piaget. The sequence essentially involved first, sorting two objects according to a common feature, then, three or more objects were sorted by a common feature, next, the concepts "some" or "all" were introduced, then, objects were classified into more than one group, and, finally, subsets and combinations of subsets were sorted.

---

[71] Matilda W. Riley et al., Sociological Studies in Scale Analysis (New Brunswick: Rutgers University Press, 1954), cited by Frederic Lord, "Scaling," Review of Educational Research, XXIV (1954), pp. 375-92.

[72] Helen C. Abell, "The Use of Scale Analysis in a Study of Differential Adoption of Homemaking Practices, Rural Sociology, XVII (1952), pp. 161-167.

[73] Ellin Kofsky, "Developmental Scalogram Analysis of Classificatory Behavior," Dissertation Abstracts, XXIV (1963), 2576.

Rater observations were employed to scale a check list for technical skills in Naval electronics.[74] The skills were ordered as to amount of inservice training required. The results indicated that the check list of technical skills was scalable. Similarly, a list of behaviors was found to be scalable by Scott in applying scalogram analysis in the investigation of delinquent behavior.[75] The list was obtained from a questionnaire covering offenses such as stealing. The respondents were asked to indicate the frequency with which they had committed each type of offense.

While Guttman had suggested its use for achievement testing, the evidence of application of the technique in this area has been fragmentary. Postove employed scalogram analysis in the development of a speechreading test.[76] She presented to adults a silent film which contained 99 conversational sentences, the subjects being required to lip read. Scalogram analysis was used to obtain 16 sentences which were reported to be scaled. The results are questionable, however, for no evidence such as a reproducibility coefficient is supplied.

---

[74] Arthur I. Siegel and Douglas G. Schultz, "Thurstone and Guttman Scaling of Job Related Technical Skills," Psychological Reports, X (1962), pp. 855-861.

[75] John Finley Scott, "Two Dimensions of Delinquent Behavior," American Sociological Review, XXIV (1959), pp. 240-243.

[76] Mary Jane Postove, "Selection of Items for a Speech-Reading Test by Means of Scalogram Analysis," Journal of Speech and Hearing Disorders, XXVII (1962), pp. 71-75.

In addition Postove used scalogram analysis to select items from an item pool, a procedure contrary to Guttman's recommendations. The resultant 16 item scale was never administered, as such, to the group. Coughenour and Christiansen developed a test of farmers knowledge of old-age and survivors' insurance.[77] The multiple choice items pertained to distinctive features of the insurance and to matters of importance for farmers' participation in the group. The test was administered as an interview, and the obtained reproducibility coefficient furnished evidence for scalability.

Neither of the above studies, however, dealt with the application of Guttman's technique to tests employed in assessing achievement of school children. The only study, to this investigator's knowledge, which related to the use of scalogram analysis with classroom achievement tests was by Bligh.[78] He applied the technique to the Paragraph Meaning, Study Skills, and Arithmetic Computation subtests of the Stanford Achievement Battery, Advanced Form J. The initial results did not warrant the acceptance of the tests as scaled; the reproducibility coefficients did not reach .80. The tests were refined by selecting items which maximized the ratios of the sums of all the covariances to the variances of the tests. The revised tests were administered to two new samples, but the obtained reproducibility coefficients still did not reach the minimum acceptable value of .90 (the range was .818 - .872). Because of the magnitude of these coefficients, however, Bligh suggested the value of further investigation of scalability in achievement testing.

---

[77] C. M. Coughenour and J. R. Christiansen, "Farmers' Knowledge: An Appraisal of Stouffer's H-Technique," Rural Sociology, XXIII (1958), pp. 253-262.

[78] Harold F. Bligh, "Empirical Investigation of Methods of Scaling Achievement Tests Based on Interelationship of Items," Dissertation Abstracts, XIX (1958), pp. 2648-2649.

To this investigator's knowledge no other studies concerning the applicability of scalogram analysis to achievement testing have been reported in the literature. With increasing demand for criterion-referenced measures comparing an individual's performance to an absolute standard independent of reference to the performance of others, the feasibility of applying this technique, in order to obtain scaled scores, should be determined. The results of a pilot study are encouraging.[79] The study involved the construction of a test in addition of whole numbers, covering concepts typically taught during early elementary education. The authors identified the objectives to be tested by, first, determining the terminal objective, then, working backwards by using as a guide the question: What skills were mastered previously in order to master this objective? A list of fifteen objectives and sample items was developed (See Appendix A).

---

[79] Glenn T. Graham and Richard C. Cox, "An Attempt to Determine the Scalability of an Elementary Math Achievement Test" (Paper read at the Pennsylvania Educational Research Association conference, Pittsburgh, Pennsylvania, April, 1965). (Mimeographed); Richard C. Cox and Glenn T. Graham, "The Development of a Sequentially Scaled Achievement Test" (Paper read at the 50th Annual Meeting of the American Educational Research Association, Chicago, Illinois, February 17, 1966). (Mimeographed.)

From two to five items were constructed for each objective. This resulted in a problem, however, for the test would undoubtedly not scale with more than one item pertaining to each objective. Rather, the test would be of the previously mentioned form discussed by Brown, et al., and Glaser, a test having a region of inconsistent responses. As a solution, each of the items corresponding to a particular objective were combined to form one "contrived item."

As an example, consider the three items:

$$\begin{array}{r}20\\ \underline{+11}\end{array}\qquad\qquad\begin{array}{r}36\\ \underline{+42}\end{array}\qquad\qquad\begin{array}{r}54\\ \underline{+33}\end{array}$$

These items would comprise one "contrived item" testing objective 8 on the list in Appendix A. Such a procedure of forming "contrived items" has been employed by Stouffer, Borgatta, Hays and Henry.[80] However, these authors formed the "contrived items" after the initial scale analysis as an aid to establishing cut-off points. In Cox and Graham's study, the "contrived items" were formed before the analysis, the cut-off points being determined exclusively by total score.

In order to obtain a substantial range of ability levels Cox and Graham administered the test to a kindergarten, first, and second grade. The students were then ranked according to total score, possible scores ranging from 0 to 15, with a contrived item considered as "passed" if two-thirds of the items comparing it were answered correctly. Inspection of the resultant response pattern indicated that some of the contrived items were not in proper order. With elimination of three contrived items, one because of its dependence on a specific curriculum and two because of ambiguous directions, and with the rearrangement of the remaining twelve contrived items a reproducibility of .977 was obtained. In order to insure against spuriously high reproducibility, Menzel's coefficient of scalability was also calculated, and equalled .902.

---

[80] Samuel A. Stouffer et al., "A Technique for Improving Cumulative Scales," Public Opinion Quarterly, XVI (1952), pp. 273-291.

In order to validate these preliminary results, the revised test was administered to different kindergarten, first, and second grade children. The analysis of their score patterns yielded a reproducibility coefficient of .970 and a coefficient of scalability of .792. The authors concluded that it was indeed possible to apply Guttman's scalogram analysis to obtain a scaled achievement test. The results, while tempered by the test's being based on a restricted area of subject matter, are encouraging for further investigation.

F. A Methodology for the Construction of a Sequentially Scaled Achievement Test

While the above study focused on the applicability of scalogram analysis to achievement testing, a methodology incorporating scalogram analysis for constructing scaled tests was concomitantly being implied. The methodology gleaned from the pilot study essentially consists of:

1. Selection of behavioral objectives in the curriculum which appear, logically, to be sequenced.

a. Identification of terminal objective.

b. Employment of question, "What skills must have been learned previously?", as a guide for selection of subsequent objectives.

2. Construction of items corresponding to each objective.

3. Combination of the items into one "contrived item."

4. Establishment of a criterion for passing each contrived item.

5. Administration and scoring of the test.

6. Application of Guttman "scalogram analysis" technique including computation of the reproducibility coefficient.

7. Computation of Menzel's "coefficient of scalability" to insure against a spuriously high reproducibility coefficient.

While successfully applied to a restricted area, further investigation of the applicability of the above methodology to a wider range of content and corresponding behavioral objectives should be attempted.

G. Evaluation of Sequentially Scaled Achievement Tests

A methodology for construction is, however, only one aspect of test development. Another important aspect of the development of such a test is the assessment of the test in terms of the typical evaluation procedures. Evaluation procedures commonly applied to standardized tests employed in the schools concern the areas of reliability, validity, and item analysis. Investigation of these evaluation procedures as they apply to scaled tests, has not, to this investigator's knowledge, been attempted. There is some evidence, however, that there are differences in evaluation procedures for norm-referenced and criterion-referenced tests, of which scaled tests are a variety. Such evidence has been reported by Cox and Vargas concerning item analysis procedures.[81]

Cox and Vargas investigated the effect of employing differential item selection techniques to identify items which discriminated according to the requirements of norm and criterion-referenced tests. For their particular criterion-referenced situation the best item would be one which was failed before training and passed afterwards. The usual norm-referenced item analysis procedures yield items which discriminate between high and low scorers after training. The authors cited an extreme example: a perfectly discriminating item for the criterion-referenced test would be one failed by all on a pretest and passed by all on a posttest. Such an item would be rejected by the norm-referenced technique at either the pretest or posttest level because it makes no discriminations among high and low scorers, being answered alike by all persons.

The authors suggested a difference index based on discriminations made between pre and posttest groups. They compared this index to the standard upper 27% - lower 27% index computed for items on each of two arithmetic tests given as pre and posttests in an individualized instruction program. Cox and Vargas indicated that if a final test consisted of the best two-thirds of the items selected by either procedure, approximately seventy-five percent to eighty percent of the items would be the same in each case. The authors noted, however, that some items not discriminating between pre and posttest groups would be retained by the upper-lower 27% procedure while some of the best discriminating items between pre and posttests groups would be discarded.

---

[81] Richard C. Cox and Julie S. Vargas, "A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests" (Paper read at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, February 17, 1966). (Mimeographed.)

While the above study was not specifically concerned with scaled tests, it did concern the area of criterion-referenced measurement which includes scaled tests. The results of the study suggest that how a test is to be employed or constructed will be a determining factor for the type of item analysis procedure required. These results support the conclusion of Husek who stated, "Unfortunately there is no evidence to demonstrate that [test] items which would be most useful for one purpose are very useful for another purpose."[82] Therefore, a test that is to be scaled may well require different item analysis procedures from a norm-referenced test. Reliability and validity may be suspect for the same reasons. Since scaled tests of achievement have not, to this investigator's knowledge, been discussed in the literature, no information regarding the characteristics of the reliability and validity of scaled tests is available. Such information should be obtained, for if the development of a scaled test is to be thorough, both construction methodology and evaluation procedures should be discussed.

---

[82] T. R. Husek, "Different Kinds of Evaluation and Their Implications for Test Development" (Paper read at the 50th Annual Meeting of the American Educational Research Association, Chicago, Illinois, February 19, 1966), p. 3. (Mimeographed.)

H. Summary

With the development of individualized instruction and similar educational innovations, criterion-referenced measures are in increased demand. With students being compared to absolute standards as criteria, what specific behaviors a student has mastered as well as how much he has mastered are desired kinds of information to be obtained from the test. Similar to norm-referenced test raw scores, criterion-referenced test raw scores have, to date, supplied most information regarding the latter (how much) and very little information regarding the former.

One solution to the problem of interpreting from a test raw score what specific behaviors a student has mastered would be a test whose items were sequentially scaled. A test so constructed would have the characteristic that an individual would pass items to a certain point. Once failing an item, he would fail all subsequent items. Therefore, a score of 4 would mean items 1, 2, 3 and 4 were passed and all other items failed.

A technique which yields tests of this type is Guttman's "scalogram analysis." While developed as a tool for attitude and opinion investigation, Guttman has suggested the use of the technique in the construction of achievement tests.

To date, however, scalogram analysis has been applied to most everything but achievement testing. In many studies the technique has yielded a sequentially scaled measuring instrument. If the same results could be obtained for achievement tests their scores would provide information indicating what specific behaviors the student has or has not mastered. An investigation of the applicability of Guttman's scalogram analysis to achievement testing is needed.

Encouraging results were obtained from a pilot study concerning the development of a sequentially scaled achievement test in the addition of whole numbers. Also from the pilot study, a methodology incorporating scalogram analysis was suggested for the construction of scaled achievement tests. Further investigation applying the methodology to a wider range of skills and objectives seems warranted.

In addition to methodology for construction, another important aspect of the test development process concerns evaluation. The evaluation procedures typically applied in the development of standardized achievement tests (norm-referenced measures) are in the areas of reliability, validity, and item analysis. With the exception of some evidence that criterion-referenced measures may require different item analysis procedures from norm-referenced measures, no evidence is available concerning the comparability of the evaluation procedures for scaled tests as opposed to standardized tests. To be thorough the development of the scaled tests should include both test construction methodology and evaluation procedures.

III. STATEMENT OF THE PROBLEM

The purpose of this study is to apply a methodology, incorporating Guttman's "scalogram analysis," for the construction of sequentially scaled achievement tests, and to develop evaluation procedures concerning the reliability, validity, and item analysis of the obtained tests.

IV. PROCEDURE

Application of the methodology for the construction of sequentially scaled tests was attempted in five areas of arithmetic achievement: addition, subtraction, numeration, time telling, and concepts in money. The behavioral objectives selected for this study pertained to skills taught in grades one through three. (See Appendix B for a listing of objectives for each test.) The sample of students was obtained from two schools in the Baldwin-Whitehall district of suburban Pittsburgh. One school, Sickman Elementary School, provided a sample of "conventional" classroom instruction; the other, Oakleaf Elementary School provided a sample of individualized instruction. All five tests were administered to both schools.

The directions for each of the items were read to the students, and ample time was provided for the student to attempt all items. The scoring criterion employed was that two-thirds of the items had to be answered correctly in order to pass a contrived item. Where only one or two items composed a contrived item, both had to be answered correctly. The scalogram analysis procedures were applied (1) to the separate test results for each school to provide evidence for test scalability for the individual schools, and (2) to the combined test results from both schools to provide an over-all indication of the tests' scalability. To take into account spuriousness in the reproducibility coefficients, Menzel coefficients of scalability were also calculated.

Following the application of the methodology for the construction of the sequentially scaled tests, evaluation was attempted in the areas of reliability, validity, and item analysis.

In considering the reliability of the scaled tests certain methods commonly employed in evaluating achievement tests were considered. One of these methods concerned the equivalence of alternate forms of a test, but this procedure was deemed inappropriate for the present investigation. The purpose of the present investigation was to attempt to develop a scaled test in each of five selected areas of arithmetic. If the methodology were successful in producing scales, then whenever alternate forms of the scales were desired, the equivalence of alternate forms procedure could be applied. Such a procedure provided no information concerning the scalability of the tests, for it is possible to obtain a high alternate forms coefficient whether a test is scaled or not.

Other reliability procedures concern the stability of a test. Typically a measure of stability, test-retest reliability, is obtained by correlating the scores from two administrations of a test, the second administration following an interval of time. The circumstances surrounding the time when the scaled tests were administered prevented any attempt to apply the test-retest procedure. Since the scaled tests had to be administered at the end of the school year to coincide with the Metropolitan Achievement Tests (this point of procedure is discussed on page 52 of this chapter), no time remained for retesting. An approximation to test-retest reliability was obtained, however, by employing the split-half reliability procedure. While providing an indication of the stability of the pupils' scores, such a procedure did contain some possible contaminating factors. According to Guilford and Ebel such a procedure functions best when the items of a test are of equal difficulty.[83] The scaled tests, however, were purposely constructed to have items of unequal difficulty. Further, because of the restricted score ranges for the scaled tests following the even-odd split, the largest range being eight, the coefficients would probably be underestimated.

In addition to the split-half procedure, stability for the scaled tests was viewed in another perspective. Rather than stability over time, a measure of the stability of the scale between groups was considered, i.e., did the test scale for more than one sample, and, if so, was the ordering of the items stable between the groups?

The rationale for determining the stability of the item orders between groups was as follows: if an order of items is established as scaled for a given group, and if that obtained scale is administered to another group for the purpose of inferring behavior from total score, the order of the items should be the same for the latter group as it was for the former. Such a measure was obtained in the present investigation. The item orders for the scaled tests administered to the Oakleaf and Sickman pupils, respectively, were obtained. These obtained orders were tested for stability with the Spearman, Rho, correlation for rank differences.

---

[83] J. P. Guilford, Fundamental Statistics in Psychology and Education (third edition; New York: McGraw-Hill Book Company, 1956), p. 456; Robert L. Ebel, Measuring Educational Achievement (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965) p. 343.

Similarly, an order of items should remain internally stable for a particular group. That is subsamples should maintain the same item orderings as the original sample if behavior inferences from total score are to be accurate. An assessment of this type of stability was also obtained. Two groups of fifty pupils each were randomly selected from each of the tests based on the combined samples from Oakleaf and Sickman. The tests of the two groups were then subjected to the scaling methodology, separately. The obtained item orders for the two groups were then tested for stability with the Spearman, Rho, correlation.

The above stability procedures differ from the traditional approach to reliability. While the traditional reliability procedures attempt to evaluate the consistency of test scores, the stability procedures attempt to evaluate consistency of item orders.

The next evaluation of the scaled tests concerned validity. Because of a multitude of types of validity, an enumeration and discussion of all types will not be presented. Rather, the present study will be limited to discussing the desired outcomes of the scaled tests and how each of these criteria were evaluated.

The scaled tests were designed to represent skills in five areas—Addition, Subtraction, Numeration, Time Telling, and Concepts in Money. Tyler stated that one criterion for validity is how clearly the objectives have been defined, and how well the items represent the objectives.[84] While validity in this sense cannot be expressed in terms of some coefficient, this criterion was employed in the formulation of the objectives and construction of the items. The objectives were defined by employing three criteria suggested by Lindvall:[85]

1. The objective should be stated in terms of the pupil.
2. The objectives should be stated in terms of observable behavior.
3. The statement of an objective should refer to the behavior or process and to the specific content to which this is to be applied.

The items were constructed directly from the obtained objectives.

The scaled tests were also designed to provide scores from which behavior could be inferred. A measure of the tests' validity in this respect would be the degree of success obtained when behavior was inferred from test score. Such a measure was provided by predicting that a total score of n for each pupil meant he had passed the first n items. Percentages were obtained for perfect predictions and predictions off by one item, and off by more than one item.

---

[84] Ralph W. Tyler, "The Development of Instruments for Assessing Educational Progress," Proceedings of the 1965 Invitational Conference on Testing Problems (Princeton: Educational Testing Service, 1966), p. 101.

[85] C. M. Lindvall, Testing and Evaluation: An Introduction. (New York: Harcourt, Brace and World, Inc., 1961), pp. 23-24.

Since the scaled tests were also designed to indicate the pupils' standing in relation to the sequence of objectives, the behaviors represented by the test score should be indicative of the students' positions in the classroom curriculum. In other words, if the test is an achievement test it should indicate how well the student is mastering the skills in the classroom. With the students at Oakleaf, a daily record was kept concerning mastery of skills in each unit of study in mathematics. A comparison of the test scores to these students' level of mastery in the respective units provides an essential measure of validity for the present study. Predictions on the basis of scaled test scores were made concerning the level and skill attained by each student in each of the five units at the time of testing. The percentages of correct predictions, predictions one, two, three, and more than three skills off were obtained.

In order to make the predictions, each behavioral objective for the five scaled tests was matched as closely as possible to a behavioral objective of the Oakleaf mathematics curriculum sequence. The Oakleaf objectives were available in a numbered sequence which was arranged by unit and level. As a result each behavioral objective from the scaled tests could be placed in a level at a skill by matching it to the Oakleaf curriculum objective. For example, the behavioral objective, "The student will be able to subtract two two-digit numerals without borrowing" was found to correspond with the Oakleaf objective at level C, skill 2 in the subtraction unit (See Appendix C for a short description of levels A-E). Since the items of the scaled test were not in the same sequential order as the objectives from the Oakleaf curriculum (see page 101), to predict the level and skill from the total score the following procedures were employed:

1. The unit and skill corresponding to the last item passed on each test were obtained for each pupil (this unit was called the "base unit").

2. If the pupil had mastered all the tested skills in the base unit, one of the following criteria were applied:

   a. If he had mastered any skills in the next unit his placement was predicted at the lowest tested skill not mastered in that unit.

   b. If he had not mastered any skills in the next unit his placement was predicted at one skill above the last tested skill in the base unit.

   c. If he had mastered all skills at the next unit criteria 1 or 2 was repeated for the subsequent unit.

   d. For those passing all items the prediction was made at one skill beyond the highest level and skill tested.

e. For those failing all items the prediction was made at one skill below the lowest level and skill tested.

3. If the pupil had not mastered all the skills in the base unit the following criteria applied:

a. If he had mastered all but one skill of the previous unit his placement was predicted at the lowest skill not mastered in the base unit.

b. If he had not mastered two or more skills in the previous unit his placement was predicted at the lowest skill not mastered in that unit.

c. If he had mastered all skills in the next unit criteria 2a, b, or c were applied.

To determine the actual level and skill for the pupils in the five units of the Oakleaf curriculum the folowing criteria were applied:

1. If the pupil worked in the unit under consideration from March on, the last skill in which he was working was used.

2. If the unit was mastered, that level was compared to the level where he was placed at the beginning of the following school year. The higher of the two was taken.

3. If the pupil had not worked in a unit since March, but did work in the unit during the year this level was compared to his placement as in (2) and the higher was taken.

4. If the pupil had not worked in the unit at all during the year, his placement for the following year was taken.

The final evaluation of the scaled test was concerned with item analysis procedures. One form of item analysis concerns the discriminating ability of the items. Traditionally items have been selected at the 50 percent level of difficulty when it was desired that items make the maximum number of discriminations between those passing and failing the items. Such procedures are not appropriate for scaled tests however, for in evaluation the concern is not always with maximum discrimination. Sometimes it is desired to know what all the students answer or what only a few can answer, or in the case of criterion-reference measures, what each student has mastered in terms of absolute standards. As Tyler states:

> Tests that are constructed to measure individual difference contain a very large proportion of items that are at the 50 percent level of difficulty because these are the most efficient in discrimination. Such tests are inappropriate for the assessment because they do not furnish an adequate picture of what is being learned by nearly all and by the most advanced.[86]

Other item analysis procedures commonly employed are the selection of items which have the highest correlation with the other items in the test, or items which have the highest correlation with the total test score. The former is especially sensitive to item difficulty, functioning best when the items are of equal difficulty.[87] Therefore, this procedure was deemed inappropriate for scaled tests, since the scaled tests were constructed to have items of unequal difficulty.

---

[86] Ralph W. Tyler, "The Development of Instruments for Assessing Educational Progress," Proceeding of the 1965 Invitational Conference on Testing Problems (Princeton: Educational Testing Service, 1966) p. 101.

[87] Guilford, *op. cit.*, p. 450.

Item-total score correlations, while not as sensitive to difficulty as item correlations, also do not afford the type of procedure necessary for scaled tests. Items which are good items for a scaled test may be judged poor by the item-total score procedure. For example, the items at the extremes of the scales which most answer correctly or incorrectly, will have low item-total score correlations and be rejected by this procedure. These items, however, may be crucial in defining a scale.

Rather than selecting items which best represent the total score, the appropriate item selection procedure for scaled tests should yield those items which best represent a scale. If each item should represent the scale then each item should have a reproducibility of .90 or greater, i.e., if there are 100 scores the maximum errors for an item should be ten. Consider, for example, a score of 4 obtained on a scaled test. Theoretically, all those individuals with a score of 4 or above should have passed item four. All those with a score of 3 or below should have failed item four. The errors for item four can be found by summing the number of people who pass item four who score less than 4 and the number of people who fail item four who score equal to or grater than 4. This procedure was employed for identifying poor items in each of the five tests for Oakleaf, Sickman, and the tests based on the combined samples.

Following the evaluation of the scaled tests, a comparison was made between the scaled tests and a norm-referenced, standardized achievement tests. Since one type of test commonly employed in the schools for pupil evaluation has been the standardized achievement tests, it was desirable to know how the results of the two types of tests compared. The Metropolitan Achievement Tests, Primary I (Form B), Primary II (Form A), and Elementary (Form C) batteries were administered to the Oakleaf pupils concurrently with the scaled tests. The comparison involved the following procedures:

1. The methods employed for the evaluation of the Metropolitan tests were applied to the scaled tests. This consisted of split-half reliability and item analysis based on a discrimination index between high and low scorers. The sizes of the reliability coefficients reported in the Metropolitan examiner's manuals were compared to the coefficients obtained from the scaled tests, taking into consideration the contaminating factors mentioned previously. The Metropolitan manuals were not specific concerning the type of discrimination index employed. Therefore, the difference between the upper twenty-seven percent and lower twenty-seven percent was chosen. The items rejected by this procedure were compared to those rejected by the suggested procedure for scaled tests.

2. An attempt was made to apply the scaling methodology to the items of the Arithmetic Computation and Concepts and Problem Solving subtests of the Metropolitan.

The items were classified with respect to the behavioral objectives they represented. If more than one item corresponded to an objective, contrived items were composed. These items and contrived items were then arranged into five tests corresponding to the five scaled tests: Addition, Subtraction, Numeration, Time Telling, and Concepts in Money. Items from the Metropolitan which did not fall into one of the above five categories were omitted. The same criteria for scalability which were applied to the scaled tests were applied to these five derived tests.

3. To determine if a particular raw score represented the same items being answered correctly, the raw score patterns for the scaled tests and the derived Metropolitan arithmetic tests were obtained. For the Metropolitan the contrived items were reduced to their corresponding individual items since the comparison involved raw scores. The raw score patterns for the five tests were compared to the raw score patterns of the scaled tests by computing the average number of score patterns per test.

The procedure for computing the average number of patterns per test is as follows:

> Perfect scores, zero scores, or scores represented by one person were omitted because only one score pattern was possible. The averages were computed by totaling the number of score patterns for a test and dividing by the smaller of the following two values: the total number of people represented by the patterns or the maximum number of possible patterns.
>
> The latter value was necessary because

some tests had so few items, and, therefore, only a small number of patterns were possible no matter how many people were represented. The maximum number of patterns was obtained by finding the combination of n things taken r at a time. For example, if a test had only three items the maximum number of patterns would be six, the scores 2 and 1 being the only scores considered.

4. The score patterns of the scaled tests and the score patterns of the derived Metropolitan scaled tests were compared as in part 3 above.

5. Predictions from the total raw scores of the derived Metropolitan tests were made concerning the level and skill of each pupil in each of the five units. Since the predictions were based on raw scores, the contrived items were again reduced to their components. With one exception the prediction followed the same procedures and criteria employed in the validation of the scaled tests. The exception involved the determination of score ranges for the Metropolitan tests. Since no contrived items were employed, many items of the Metropolitan tests pertained to the same objective. Therefore, a range of scores, corresponding to the number of items testing a given objective, was established for the purpose of predicting the pupils' positions in the curriculum. For example, if the first objective was tested with three items a score of 1, 2, or 3 would correspond to the same level and skill. If the next objective had two items the scores of 4 and 5 would correspond to the next

skill. Percentages were obtained in the same manner as the scaled tests. These percentages were compared to those of the scaled tests to determine the relative success of each type of raw score.

A. Summary

The methodology for the construction of sequentially scaled achievement tests was applied in the development of tests in five areas of arithmetic achievement: Addition, Subtraction, Numeration, Time Telling, and Concepts in Money. Subjects were obtained from two elementary schools, Oakleaf and Sickman, in suburban Pittsburgh. Following the construction of the tests, three phases of evaluation were attempted--reliability, validity, and item analysis. The following procedures were employed:

1. For reliability (a) the split-half (even-odd) correlation was obtained as an approximation to test-retest reliability, and (b) Spearman Rho rank order correlations were obtained for the orderings of two random samples of pupils selected from the combined Oakleaf and Sickman populations. Rank order correlations were also obtained for the item orderings of the Oakleaf tests versus the corresponding Sickman tests. These provided a measure of the stability of the items in the scales.

2. The essential measure of validity was provided by the agreement between the scaled scores and the Oakleaf student's level of mastery in each of the five units. Predictions were made, on the basis of the scaled test scores,

concerning where each student should be in each of the five units at the time of testing. The correspondence of these predictions to the daily record of each student was reported by percentage perfect predictions, predictions off by one, two, three, and more skills.

A measure of the validity of each scaled test was also obtained by computing the percentage of times a total score of n represented a student's passing the first n items. Perfect representations, and those one item off, or more than one item off were obtained.

3. The item analysis procedure involved establishing a minimum reproducibility of .90 for each item. The maximum number of errors for an item was ten percent of the number of people. Errors were obtained by counting those persons having a score lower than the item number but passing the item, and those persons obtaining a score equal to or greater than the item number but failing the item.

The final phase of the present investigation involved a comparison of the scaled tests results to the results of the Metropolitan Achievement Tests administered concurrently to the Oakleaf pupils. The comparison involved the following procedures:

1. The methods employed for the evaluation of the Metropolitan Tests were applied to the scaled tests. Reliability and item analysis procedures were compared.

2. An attempt was made to determine whether the scaling methodology could be applied to the items of the arithmetic computation subtest of the Metropolitan

Achievement Test.

3. To determine if a particular raw score represented the same items being answered correctly the score patterns for the scaled tests and the raw scores and scaled scores of the derived Metropolitan were compared by number of patterns for individual and number of patterns for test.

4. The raw scores of the five tests derived from the Metropolitan were used to infer behavior from total score. The predictions and percentages were obtained in the same manner as the scaled tests. The resultant percentages provided a measure of the relative success of both tests.

V. RESULTS AND DISCUSSION

A. Application of Scaling Methodology

The scaling methodology for the present investigation consisted of the following steps:

1. Identification of the terminal behavioral objective to be tested, followed by identification of a series of behavioral objectives which appear logically to preceed the terminal objective in a sequence.
2. Construction of items corresponding to each objective.
3. Combination of the items into "contrived items."
4. Establishment of a criterion for passing each "contrived item."
5. Administration and scoring of the tests.
6. Application of Guttman's "scalogram analysis" technique including computation of the reproducibility coefficient.
7. Computation of Menzel's coefficient of scalability.

For both the Oakleaf and Sickman tests an initial reproducibility coefficient was calculated. The items were then rearranged, where necessary, to obtain the maximum coefficient of reproducibility. The coefficient of scalability was computed following the final arrangement of the items (See Appendix D for a Sample Scalogram). The obtained reproducibility and scalability coefficients are presented in Table 1 and Table 2.

TABLE 1

REPRODUCIBILITY AND SCALABILITY COEFFICIENTS FOR OAKLEAF SCALES

	Repro. I	Repro. II	Scalability	Revised Repro.	Revised Scal.
Subtraction	.910	.966	.815	.961	.791
Addition	.961	.979	.854	.977	.825
Numeration	.957	.965	.682	.947	.647
Money	.943	.953	.690	.951	.676
Time	.907	.929	.711	.920	.676

TABLE 2

REPRODUCIBILITY AND SCALABILITY COEFFICIENTS FOR SICKMAN SCALES

	Repro. I	Repro. II	Scalability	Revised Repro.	Revised Scal.
Subtraction	.921	.967	.822	.958	.822
Addition	.958	.966	.797	.951	.729
Numeration	.948	.968	.799	.952	.799
Money	.952	.968	.801	.964	.786
Time	.918	.918	.675	.909	.644

It will be noted that both tables contain columns for revised reproducibility and revised scalability coefficients. These revisions were necessary because of the number of perfect scores occurring. With a perfect score the reproducibility has to be perfect and, therefore, a group of perfect scores will spuriously raise the reproducibility of the test as a whole. While the coefficient os scalability will insure against spuriousness in the test, it too must be perfect for all perfect scores. Therefore, all but one perfect paper was omitted from each test, the one perfect score remaining to give the test a ceiling. Reproducibility and scalability coefficients were again computed, and these coefficients constituted the revised coefficients appearing in the tables. These revised coefficients were established as the criteria for accepting or rejecting the tests as scales.

From the values presented in Tables 1 and 2, four of the five Oakleaf tests and four of the five Sickman tests met the criteria for being a scale. Because the scalability coefficient was below .65 the Oakleaf Numeration test failed to meet the criteria and was not considered scaled for this group. The Oakleaf Money and Time tests barely reached the minimum value for scales, and while accepted as scaled, were in need of further revision. The omission of some poor items (to be discussed later) and/or the addition of new items is suggested.

The Sickman Time test also failed to meet the scalability criterion and was not considered scaled for the Sickman pupils. The remaining four tests had substantial scalability coefficients in contrast to only two substantial scalability coefficients for the Oakleaf tests. The difference between the Oakleaf and Sickman Numeration tests may be a function of differential familiarity to Numeration content. The lowest score on the Oakleaf Numeration test was 7 while there were nineteen scores below 7 for the Sickman Numeration test. It would appear that the Oakleaf students had greater familiarity with the Numeration objectives tested, the test being too easy for them. As a result the objectives and corresponding items were not sequenced once the students had become familiar with them. The Sickman pupils who scored lower than 7 evidently had not had the same amount of familiarity with the Numeration objectives, and could not attempt the higher level items. Consequently, the opportunity to have high reproducibility and scalability was greater for these items.

The same conjecture could not apply to the Money test, however. The score ranges were almost the same for both schools. Perhaps the best suggestion is that the Money test functions differently for the two groups, and that it best represents the ordering of the objectives of the Sickman curriculum.

Following the attempt to scale the tests for each school separately, the two samples of pupils were combined to determine if the tests would scale for the entire group. The reproducibility and scalability coefficients are reported in Table 3 for the final ordering of items. Revised reproducibility and scalability coefficients were again calculated by omitting all but one perfect paper from each test. These revised coefficients were employed as the criteria for considering a test as scaled. The results show that only the Time test failed to meet the criteria and was not considered scaled. The magnitude of the scalability coefficients for the Numeration and Money tests was undoubtedly enhanced by the Sickman scores.

TABLE 3

REPRODUCIBILITY AND SCALABILITY COEFFICIENT
FOR THE COMBINED SAMPLE OF PUPILS

	Repro. I	Scalability	Revised Repro.	Revised Scal.
Subtraction	.955	.778	.946	.739
Addition	.973	.837	.966	.776
Numeration	.962	.715	.943	.715
Money	.958	.733	.955	.712
Time	.917	.669	.907	.630

B. Reliability of the Scaled Tests

Having applied the methodology for the construction of scaled tests, the resultant five tests were subjected to evaluation procedures in the areas of reliability, validity, and item analysis. The initial measure of reliability involved the computation of split-half reliability coefficients to afford an indication of the stability of the test scores.

Because the final orders of items of each test were arranged in ascending order of difficulty an even-odd split was chosen. To afford an adequate sample size the correlations were computed from the combined Oakleaf and Sickman test results. The correlations, appearing in Table 4, were corrected by using the Spearman-Brown formula.

TABLE 4

CORRECTED SPLIT-HALF RELIABILITY COEFFICIENTS FOR THE SCALED TESTS

Test	r_{tt}
Subtraction	.872
Addition	.908
Numeration	.931
Money	.852
Time	.787

From the magnitudes of the split-half coefficients the tests appear to have fairly stable scores. The Time test had the lowest coefficient, a result which indicated some lack of stability but a result not unexpected since the Time test did not meet the scale criteria. When the contaminating factors of the split-half correlation are considered, the above results are most encouraging, for with varying difficulties and restricted score ranges the coefficients may be underestimated.

While the split-half reliability afforded a measure of the stability of the scores, a measure of the stability of the item orderings was also computed. Two assessments of the stability of the tests' item orders were employed: (1) the stability of the item orderings between the two schools and (2) the stability of the item orderings for two randomly selected groups from the combined samples. The Spearman Rho, rank-difference correlation was employed for both assessments.

The Spearman correlation coefficients for the item orders between schools appear in Table 5. The Subtraction and Addition test items had stable item orders. The Time and Money item orders were not as stable and indicated some fluctuation. A coefficient as large as .880 was unexpected for the Time test. The order of items remained relatively stable even though the test did not scale. The Numeration item orders were not stable. This result, however, reinforces the conjecture concerning this test's scalability; the items did not have a scalable ordering for Oakleaf but did for Sickman where they were more difficult.

TABLE 5

SPEARMAN RHO CORRELATION COEFFICIENTS FOR OAKLEAF AND SICKMAN ITEM ORDERS

Test	Rho
Subtraction	.946
Addition	.982
Numeration	.736
Money	.846
Time	.880

A second assessment of the stability of the items was made by selecting two random samples of fifty pupils from the tests based on the combined samples of Oakleaf and Sickman. Spearman Rho correlations were obtained for the two samples and are presented in Table 6. The results were that the item orderings are stable within any of the five combined tests.

TABLE 6

SPEARMAN RHO CORRELATION COEFFICIENTS FOR ITEM ORDERS OF TWO RANDOM SAMPLES FROM THE TESTS BASED ON COMBINED SAMPLES

Test	Rho
Subtraction	.964
Addition	.982
Numeration	.930

TABLE 6 (continued)

SPEARMAN RHO CORRELATION COEFFICIENTS FOR ITEM ORDERS OF TWO RANDOM SAMPLES FROM THE TESTS BASED ON COMBINED SAMPLES

Test	Rho
Money	.930
Time	.932

C. Validity of the Scaled Tests

The second phase in the evaluation of the scaled tests concerned validity. One of the validation procedures was the measure of the degree of success obtained when behavior was inferred from the total raw score of the scaled tests. Theoretically, a pupil obtaining a score of n should have answered only the first n items correctly. Percentages for each test were obtained for the number of times the total score equaled the first n items passed. Percentages were also obtained for the number of times the total score was off by one item or more than one item in representing the first n items passed. Perfect scores were omitted from the calculations since those students had not reached a point at which an item was failed. The resultant percentages for Oakleaf are presented in Table 7. The Sickman results are presented in Table 8. After the two schools' samples were combined, the item orders were reanalyzed. The results for the combined samples are presented in Table 9. The number in parentheses following the title indicates the number of scores employed in the calculations.

TABLE 7

PERCENTAGES OF RAW SCORES EQUALING THE FIRST n ITEMS PASSED (OAKLEAF)

Test	Percent Perfect	Percent One Item Off	Percent Greater Than One Item Off
Subtraction (77)	59.7	37.7	2.6
Addition (72)	70.8	25.0	4.2
Numeration (54)	33.3	57.4	9.3
Money (79)	54.4	32.9	12.7
Time (74)	17.6	45.9	36.5

TABLE 8

PERCENTAGES OF RAW SCORES EQUALING THE FIRST n ITEMS PASSED (SICKMAN)

Test	Percent Perfect	Percent One Item Off	Percent Greater Than One Item Off
Subtraction (57)	61.4	31.6	7.0
Addition (50)	40.0	50.0	10.0
Numeration (47)	42.6	46.8	10.6
Money (67)	64.3	28.3	7.4
Time (70)	17.1	34.3	48.6

TABLE 9

PERCENTAGES OF RAW SCORES EQUALING THE FIRST n ITEMS PASSED (COMBINED)

Test	Percent Perfect	Percent One Item Off	Percent Greater Than One Item Off
Subtraction (134)	45.5	50.0	4.5
Addition (122)	57.4	36.9	5.7
Numeration (101)	30.7	58.4	10.9
Money (146)	56.8	34.2	8.9
Time (144)	14.6	32.6	52.8

The results showed that for every test but Time the total scores were within a maximum of one item off in having the total represent the first n items passed for approximately ninety percent of the cases. These percentages are closely related to the reproducibility of the tests, since the errors employed in determining the reproducibility coefficients were the same errors made when the total score did not equal the first n items passed.

The Time test, having the lowest reproducibility in all cases was also the poorest in this phase of validity evaluation. The scores did not come within one item of equaling the first n items passed in more than fifty percent of the cases for Sickman, and more than sixty percent for Oakleaf. When the schools were combined the Time scores were successful in being within one item in just slightly more than fifty percent of the cases.

The final assessment of the validity of the scaled tests, and perhaps the most crucial for the present investigation, was how well the tests' results indicated the position of the pupils in the classroom curriculum sequence. In order to determine this type of validity a daily record of the pupils' positions by skill and level in each unit was required. Such a record was maintained for the Oakleaf pupils but not for the Sickman pupils. The results, therefore, pertained only to the Oakleaf tests. On the basis of raw score alone, and following the procedures outlined in the previous chapter predictions were made for each pupil concerning the level and skill at which he should be working in each of the five tested areas.

Percentages were obtained for the number of times the pupil was placed at the exact level and skill or at one, two, three or more skills off. The obtained percentages are shown in Table 10. The number in parentheses beside each test name indicates the number of pupils who could be place with a degree of accuracy. In a few instances pupils were no longer in the school and no placement data for 1965 was available.

TABLE 10

PERCENTAGES OF PREDICTIONS OF PUPIL POSITION IN THE OAKLEAF CURRICULUM SEQUENCE ON THE BASIS OF SCALED TEST RAW SCORE

Test	Percent Perfect	Percent One Skill Off	Percent Two Skills Off	Percent Three Skills Off	Percent More Than Three Skills Off
Subtraction (88)	28.4	26.1	26.1	12.5	6.8
Addition (82)	2.4	3.7	18.3	20.7	54.9
Numeration (84)	34.5	33.3	20.2	0	11.9
Money (82)	6.1	25.6	14.6	35.4	18.3
Time (80)	1.2	41.2	13.8	23.8	20.0

The results indicated that in four of the five tests at least eighty percent of the Oakleaf pupils were placed within a maximum of three skills from their position in the curriculum sequence. Approximately one third to two thirds of the pupils, depending on the test, were placed at a maximum of one skill off.

The validity of the Addition test was poor, especially in comparison to the other tests. Since the Addition test had ranked so highly with respect to the other evaluation criteria, the result was surprising. The result should have been expected, however, when viewed with the structure of the curriculum sequences for the five tested areas. The information provided in Table 11 indicated the number of skills in the Oakleaf curriculum at each of the five levels tested in each unit. Table 12 furnished the ratios of the number of skills in the last three levels (C, D, and E) to the number of items in each respective test. The reason for eliminating levels A and B was that when the predictions were made to determine the validity of the tests, only twenty-one of the 416 cases were actually in level B and none were in level A.

TABLE 11

NUMBER OF CURRICULUM SKILLS PER LEVEL FOR TESTED UNITS

Tested Units	Levels A	B	C	D	E	Total
Subtraction	2	6	2	3	3	16
Addition	2	8	6	5	7	28
Numeration	10	9	4	2	3	28
Money	1	3	2	5	3	14
Time	3	4	7	3	2	19

TABLE 12

RATIOS OF NUMBER OF TEST ITEMS TO NUMBER OF CURRICULUM SKILLS IN LEVELS C,D, and E

Tests	Total Skills C + D + E	Number of Test Items	Items/Skills
Subtraction	8	11	1.38
Addition	18	14	.78
Numeration	9	14	1.56
Money	10	12	1.20
Time	12	16	1.33

From the results of these two tables the Addition test should have been the least valid, it employed slightly greater than half the number of items as the other tests to evaluate a skill. In other words a smaller number of items were employed to test a wider range of skills. While, on the average, each skill was tested with an item on each of the other tests, more than one skill was tested with an item on the Addition test.

D. Item Analysis of the Scaled Tests

The final evaluation procedure suggested for the scaled tests involved item analysis. The scalogram analysis procedure actually included an item analysis technique. Since the minimum reproducibility for a test was .90, the criterion adopted for the analysis of the items was that each item should have a minimum reproducibility of .90. Rather than computing reproducibility coefficients for all the items the maximum number of errors for an item were found. An error was counted when (1) a correct response was made by an individual to an item whose number was greater than the total score obtained by that individual, and (2) an incorrect response was made to an item whose number was lower than the total score of the individual. Any item whose errors exceeded the maximum number of errors was considered as a poor item. The procedure was applied to the Oakleaf, Sickman, and combined tests. The results are shown in Tables 13, 14, and 15. The number in parentheses after each test name indicates the maximum number of errors for the items of that test. An item number superscribed with a bar signifies a poor item.

TABLE 13

ITEM ANALYSIS FOR OAKLEAF SCALED TESTS

Subtraction (8)																
Item No.	1	2	3	4	5	$\overline{6}$*	7	8	$\overline{9}$	10	11					
Errors	2	1	7	8	4	12	4	6	10	8	4					
Addition (8)																
Item No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14		
Errors	0	0	1	0	2	4	4	4	5	7	8	5	5	3		
Numeration																
Item No.	1	2	3	4	5	6	7	8	9	10	$\overline{11}$	$\overline{12}$	$\overline{13}$	$\overline{14}$		
Errors	0	0	1	3	3	4	4	3	6	8	10	10	10	20		
Money (8)																
Item No.	1	2	3	4	5	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$	$\overline{10}$	11	12				
Errors	1	4	6	5	6	9	15	19	11	14	4	0				
Time (8)																
Item No.	1	2	3	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{7}$	8	9	$\overline{10}$	$\overline{11}$	$\overline{12}$	$\overline{13}$	$\overline{14}$	$\overline{15}$	$\overline{16}$
Errors	1	3	8	13	17	17	15	5	9	18	15	25	12	10	12	12

*An item number superscribed with a bar signifies a poor item.

TABLE 14

ITEM ANALYSIS FOR SICKMAN SCALED TESTS

Subtraction (7)

Item No.	1	2	3	4	5	6	7	8	9	10	$\overline{11}$*
Errors	2	5	7	4	5	6	3	3	7	4	8

Addition (7)

Item No.	1	2	3	4	5	6	7	8	9	10	$\overline{11}$	12	13	14
Errors	0	0	0	4	3	5	6	6	6	7	19	1	5	7

Numeration (7)

Item No.	1	2	3	4	5	$\overline{6}$	7	8	9	10	11	12	13	14
Errors	1	6	7	2	4	9	3	5	6	4	5	1	3	3

Money (7)

Item No.	1	2	3	4	5	6	7	$\overline{8}$	$\overline{9}$	$\overline{10}$	11	12
Errors	1	1	3	7	2	7	6	8	9	9	5	0

Time (7)

Item No.	1	2	3	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$	$\overline{10}$	$\overline{11}$	$\overline{12}$	$\overline{13}$	14	$\overline{15}$	16
Errors	3	4	6	11	12	17	15	12	13	25	30	28	11	6	8	5

*An item number superscribed with a bar signifies a poor item.

TABLE 15

ITEM ANALYSIS FOR COMBINED SCALED TESTS

Subtraction (16)

Item No.	1	2	3	4	$\overline{5}$	6	7	8	$\overline{9}$	$\overline{10}$	$\overline{11}$
Errors	4	6	14	13	17	13	7	9	38	35	18

Addition (15)

Item No.	1	2	3	4	5	6	7	8	9	10	$\overline{11}$	12	13	14
Errors	0	0	1	4	5	11	10	10	10	13	27	6	11	10

Numeration (15)

Item No.	1	2	3	$\overline{4}$	5	6	7	8	9	$\overline{10}$	11	12	13	$\overline{14}$
Errors	0	9	8	17	4	12	15	10	10	16	14	12	14	23

Money (15)

Item No.	1	2	3	4	5	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$	$\overline{10}$	11	12
Errors	3	10	9	7	9	19	19	21	31	23	9	0

Time (16)

Item No.	1	2	3	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$	$\overline{10}$	$\overline{11}$	$\overline{12}$	$\overline{13}$	14	$\overline{15}$	$\overline{16}$
Errors	5	6	14	26	28	29	36	41	33	43	48	47	23	16	17	18

Only the Oakleaf Addition test contained no poor items for this analysis. The Addition test as a whole had good items, only one item was poor for both the Sickman and combined tests. The Sickman Subtraction and Numeration tests were the only other tests to have as few as one poor item. The remaining tests had several poor items, but the Time test was by far the worst for all situations, the poor items outnumbered the good items almost three to one.

To determine if the poor items were the same or variable for the Oakleaf, Sickman and combined tests the items were renumbered according to their original order--the order in which they appeared when the tests were administered. The results are presented in Table 16.

TABLE 16

ORIGINAL ITEM ORDERS FOR POOR ITEMS

Tests	Original Order
Subtraction Oakleaf	5 6
Subtraction Sickman	6
Subtraction Combined	5 6 7 11
Addition Oakleaf	No poor items
Addition Sickman	9
Addition Combined	9
Numeration Oakleaf	3 12 13 14
Numeration Sickman	8
Numeration Combined	5 11 12
Money Oakleaf	5 7 8 9 10
Money Sickman	5 8 10
Money Combined	5 7 8 9 10
Time Oakleaf	4 5 6 7 8 9 10 11 13 14 15 16
Time Sickman	4 5 6 7 8 9 10 11 12 13 15
Time Combined	4 5 6 7 8 9 10 11 12 13 15 16

With the exception of the Numeration test the same poor items were generally present in at least two out of the three test situations. There was variation, however, between the schools, for the poor items were never all alike for any of the tests.

The Numeration test represented the greatest amount of variability with seven of eight poor items being different for the three test situations. The results also showed that four items were poor for Oakleaf and only one item was poor for Sickman. This affords further information concerning why the Sickman Numeration test was scaled and the Oakleaf test was not.

E. Comparison of Scaled Tests and Metropolitan Achievement Tests

Following the evaluation of the scaled tests, certain comparisons were made between these tests and the *Metropolitan Achievement Tests*. The comparisons were designed to investigate similarities and differences in the evaluation procedures and results of the two types of tests. The comparisons included the following:

1. The initial comparison was an attempt to apply to the scaled tests the procedures employed in the evaluation of the Metropolitan tests. The examiner's manual of the Metropolitan described two evaluation procedures: (a) split-half reliability and (b) item analysis involving a discrimination index between high and low scorers. The split-half coefficients for the scaled tests, obtained previously, are presented in Table 17 along with the coefficients for the arithmetic subtests of the Metropolitan test batteries. The latter were taken from the examiner's manuals and are expressed as ranges of coefficients.

TABLE 17

SPLIT-HALF RELIABILITY COEFFICIENTS FOR SCALED TESTS AND METROPOLITAN ACHIEVEMENT TESTS

Test	r_{tt}
Scale Subtraction	.872
Scale Addition	.908
Scale Numeration	.931
Scale Money	.852
Scale Time	.787
Metro. I Arith. Concepts	.81 - .89
Metro. I Arith. Skills	.94 - .95
Metro. II Arith. Concepts	.80 - .87
Metro. II Computation	.74 - .88
Metro. III Arith. Concepts	.86 - .91
Metro. III Arith. Computation	.91 - .93

The coefficient values for both types of tests are similar. Of the two contaminating factors for split-half coefficients, only the range of difficulty was a factor for the Metropolitan, and this factor undoubtedly not to the fullest extent since easy items for all pupils and difficult items for all pupils would have been rejected in the item analysis. Therefore, the scaled test coefficients may be underestimated to a greater extent than the Metropolitan coefficients.

The second evaluation procedure, item analysis, was compared by applying the method used in the evaluation of the Metropolitan to the Oakleaf scaled tests. The manuals for the Metropolitan were not specific concerning the particular high-low discrimination index employed nor the criteria established for rejection of an item. Therefore, percent difference was obtained and items with a difference of thirty percent or less were rejected. The item analysis is presented in Table 18. Items superscribed with bars are poor items. The number in parentheses after the test name indicates the number of persons in the upper and lower groups, respectively. The similarity of this item analysis procedure to the scaled test item analysis procedure can be obtained from Table 19.

TABLE 18

UPPER 27 PERCENT—LOWER 27 PERCENT ITEM ANALYSIS FOR OAKLEAF SCALED TESTS

Subtraction (24)

Item No.	$\overline{1}$*	$\overline{2}$	$\overline{3}$	4	5	6	7	8	9	10	11
Upper Percent	100	100	100	100	100	100	100	100	100	100	46
Lower Percent	92	92	71	63	50	63	25	17	04	0	0
Difference	08	08	29	37	50	37	75	83	96	100	46

Addition (23)

Item No.	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{6}$	7	8	9	10	11	12	13	14
Upper Percent	100	100	100	100	100	100	100	100	100	100	96	100	91	57
Lower Percent	100	100	91	87	78	70	61	52	52	43	13	09	0	0
Difference	0	0	09	13	22	30	39	48	48	57	83	91	91	57

Numeration (23)

Item No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Upper Percent	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Lower Percent	100	100	100	91	96	87	83	83	74	74	78	22	17	04
Difference	0	0	0	09	04	13	17	17	26	26	22	78	83	96

TABLE 18 (continued)

UPPER 27 PERCENT—LOWER 27 PERCENT ITEM ANALYSIS FOR OAKLEAF SCALED TESTS

Money (23)

Item No.	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	5	6	7	8	9	10	11	$\overline{12}$
Upper Percent	100	100	100	100	100	100	100	100	100	100	100	17
Lower Percent	96	83	75	70	65	61	57	48	30	13	0	0
Difference	04	17	26	30	35	39	43	52	70	87	100	17

Time (23)

Item No.	$\overline{1}$	$\overline{2}$	$\overline{3}$	4	$\overline{5}$	6	7	8	9	10	11	12	13	14	15	16
Upper Percent	100	100	100	100	100	100	100	100	100	96	96	91	96	100	70	78
Lower Percent	96	100	83	57	91	61	26	17	17	35	17	30	04	0	0	0
Difference	04	0	17	43	09	39	74	83	83	61	79	61	92	100	70	78

*An item number superscribed with a bar signifies a poor item.

TABLE 19

ITEM DESIGNATED AS POOR BY TWO ITEM ANALYSIS PROCEDURES: THE SCALE PROCEDURE AND UPPER-LOWER 27 PERCENT

	Subtraction Item Numbers
Scale	6, 9
Upper-Lower	1, 2, 3
	Addition Item Numbers
Scale	None
Upper-Lower	1, 2, 3, 4, 5, 6
	Numeration Item Numbers
Scale	11, 12, 13, 14
Upper-Lower	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
	Money Item Numbers
Scale	6, 7, 8, 9, 10
Upper-Lower	1, 2, 3, 4
	Time Item Numbers
Scale	1, 2, 3, 5
Upper-Lower	4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16

These results indicated that the items rejected by the two procedures were almost completely different. In each test the upper-lower method rejected the first several items. These items were easy items for most of the pupils because they tested behavioral objectives appearing early in the curriculum sequence. While the upper-lower item analysis rejected these easy items because they did not discriminate among the high and low scores, the scaled test method did not reject these items because they were highly reproducible and provided an assessment of the pupils' mastery at the lower levels. Items chosen for a scale start with the very easy items and progress to the most difficult. Each of the items selected by the scaled test item analysis should best constitute a scale. Each of the items chosen by the upper-lower twenty-seven percent item analysis should discriminate between high and low scorers. The type of item which would be rejected by both procedures would be the item answered correctly by low scorers and incorrectly by high scorers.

2. The second comparison of the scaled tests and the Metropolitan tests was an attempt to apply the scaling methodology in order to derive scaled tests from the items of the Metropolitan arithmetic subtests. All the items of the arithmetic subtests were classified according to their respective behavioral objectives. In some cases more than one item tested a particular objective. When this occurred, contrived items were constructed, and the criterion of two-thirds correct for passing was employed. The items were arranged into tests corresponding to the five areas covered by the scaled tests. No items were included which did not represent one of the five areas.

The Primary I Metropolitan test yielded only one test Numeration, which had sufficient items (eight) to apply the scale criteria. Guttman had suggested a minimum of ten items, but the criterion was extended to see what results would be obtained. Of the remaining for tests from the Primary I battery, Subtraction, Addition, and Money had three items each. Time was tested with one item.

The Primary II battery yielded three tests with sufficient items to attempt the scale criteria, Addition, seven items; Subtraction, eight items; Numeration, eight items. Money was tested with four items and Time with three, both insufficient. The Elementary battery yielded two tests with sufficient items, Addition, fourteen; Subtraction, nine. Numeration had three items; Money, five; and Time, one. The resultant coefficients are shown in Table 20. Only revised coefficients were calculated. The number in parentheses after the test name indicates the number of scores employed in the calculations.

TABLE 20

REPRODUCIBILITY AND SCALABILITY OF TESTS DERIVED FROM THE METROPOLITAN ACHIEVEMENT TESTS

Test	Revised Reproducibility	Revised Scalability
Primary I Numeration (17)	.934	.550
Primary II Addition (20)	.986	.780
Primary II Subtraction (27)	.921	.605
Primary II Numeration (11)	.920	.364
Elementary Addition (27)	.926	.636
Elementary Subtraction (29)	.958	.744

Only two of the tests met the criteria for scalability, Primary II Addition and Elementary Subtraction, but neither would have been subjected to the scaling criteria had the minimum number of items requirement not been extended. The only test for the areas investigated which had the minimum number of items was the Elementary Addition test, but it failed to meet the scale criteria. One other test, which was not included in the five areas, had sufficient items for scaling. This was a multiplication test in the Elementary battery. There were twelve items for this test but the scalability coefficient was too low, .629. The reproducibility coefficient was acceptable at .925.

3. The third comparison between the two types of tests concerned the number of score patterns obtained for a

given test score in order to determine if a particular raw score represented the same items being answered correctly. The initial phase of this comparison involved the raw scores of the scaled tests and the raw scores of the tests derived from the Metropolitan batteries. For each test the ratio of the number of score patterns per test was obtained. This ratio employed one of the following two denominators, whichever was the minimum: (1) the total number of subjects taking the test or (2) the maximum total number of possible score patterns. Perfect scores, zero scores, and scores obtained by only one individual were omitted from the analysis because only one score pattern was possible for each. Since the fewest number of patterns should be present to best interpret the items represented by a score, the ratio of patterns to test should be as small as possible. The ratios are represented in Table 21. For Time only the Metropolitan II test was included; the Primary I and Elementary Metropolitan tests had only one item each for time. The table includes only the maximum values employed. When total subjects is the smaller it is presented, when total patterns is the smaller only that value is presented.

TABLE 21

ANALYSIS OF NUMBER OF RAW SCORE PATTERNS PER TEST

Test	Total Subjects	Total Maximum Patterns	Total Patterns	Patterns/Test
Scale Subt.	76		30	.39
Scale Add.	71		26	.37
Scale Num.	53		27	.51
Scale Money	78		32	.41
Scale Time	73		62	.85
Metro. I Subt.	19		13	.47
Metro. II Subt.	24		10	.92
Metro. III Subt.	28		22	.64
Metro. I Add.	17		13	.76
Metro. II Add.	19		10	.53
Metro. III Add.	22		22	1.00
Metro. I Num.	19		11	.58
Metro. II Num.	6		3	.50
Metro. III Num.	24		11	.46
Metro. I Money		14	9	.64
Metro. II Money	21		9	.43
Metro. III Money	26		3	.12
Metro. II Time		14	6	.43

The results appeared inconsistent at first. The patterns per test ratios for the scaled Addition and Subtraction tests were all lower and, therefore, superior to the corresponding Metropolitan tests. The Numeration tests were relatively the same, the scaled Money test was superior to the Metropolitan I test, about equal to the Metropolitan II test, and inferior to the Metropolitan III test. The scaled Time test was inferior to the Metropolitan II Time test. It should be remembered, however, that only the Addition and Subtraction scaled tests had good scalability and reproducibility coefficients for Oakleaf. The Time and Money tests barely met the criteria and the Numeration test did not meet the criteria. Therefore, only the Addition and Subtraction tests were good examples of scaled tests in the present situation.

The second phase in the comparison of score patterns concerned the scaled tests and the tests which were derived to be scaled from the Metropolitan batteries. While only two of the Metropolitan tests met the criteria for scaling and many did not have enough items to even attempt the calculation of the criteria, all were included in this analysis. The reason was to determine if any of the tests had improved pattern per test ratios which were comparable to the scaled tests. The results are presented in Table 22. Again, only those maximum values employed in the calculation are presented in the table.

TABLE 22

ANALYSIS OF SCORE PATTERNS PER TEST FOR SCALES DERIVED FROM THE METROPOLITAN TESTS

Test	Total People	Total Maximum Patterns	Total Patterns	Patterns/Test
Metro. I Subt.		6	2	.33
Metro. II Subt.	29		22	.76
Metro. III Subt.	27		12	.44
Metro. I Add.		3	1	.33
Metro. II Add.	19		4	.21
Metro. III Add.	24		23	.96
Metro. I Num.	15		5	.33
Metro. II Num.	9		5	.56
Metro. III Num.		6	5	.83
Metro. I Money		6	4	.67
Metro. II Money		14	7	.50
Metro. III Money	26		3	.12
Metro. II Time		6	4	.67

Comparing the results of Table 22 with those of Table 21, the scaling methodology resulted in improved patterns per test ratios for the derived Metropolitan Addition and Subtraction tests, and in some cases the ratios were superior to those of the scaled tests. The remaining results were inconsistent, improving in some tests, and showing greater patterns per test ratios in other cases.

4. The final comparison between the scaled tests and the Metropolitan tests involved the validity of the tests in terms of how accurately they predicted the pupils positions in the five units of the Oakleaf curriculum sequence. The predictions followed the same criteria established previously for the scaled tests, and were based on raw scores. As a result, no contrived items were present for the Metropolitan tests. The results were compared by grade, and only those pupils who had taken both tests were included. The results, presented in Table 23, indicate the percentages of perfect predictions and predictions one, two, three, and more than three skills off. The number in parentheses after the test name is the number of predictions employed in the calculations.

TABLE 23

PERCENTAGES OF PREDICTIONS OF PUPIL POSITION IN THE OAKLEAF CURRICULUM SEQUENCE: SCALED TEST RAW SCORE AND METROPOLITAN TEST RAW SCORE

Test and Grade	Percent Perfect	Percent One Skill Off	Percent Two Skills Off	Percent Three Skills Off	Percent More Than Three Skills Off
Scale Subt. 1st (30)	37	03	13	30	17
Metro. Subt. 1st (30)	20	43	10	07	20
Scale Subt. 2nd (27)	26	22	44	04	04
Metro. Subt. 2nd (27)	10	59	0	15	15
Scale Subt. 3rd (28)	25	50	22	04	0
Metro. Subt. 3rd (28)	36	50	07	0	07
Scale Add. 1st (30)	03	07	10	27	53
Metro. Add. 1st (30)	0	0	23	13	63
Scale Add. 2nd (23)	0	04	17	30	48
Metro. Add. 2nd (23)	13	48	0	13	26
Scale Add. 3rd (27)	04	0	26	07	63
Metro. Add. 3rd (27)	04	22	33	26	15

TABLE 23 (continued)

PERCENTAGES OF PREDICTIONS OF PUPIL POSITION IN THE OAKLEAF CURRICULUM SEQUENCE: SCALED TEST RAW SCORE AND METROPOLITAN TEST RAW SCORE

Test and Grade	Percent Perfect	Percent One Skill Off	Percent Two Skills Off	Percent Three Skills Off	Percent More Than Three Skills Off
Scale Num. 1st (30)	57	20	17	0	06
Metro. Num. 1st (30)	03	0	0	10	87
Scale Num. 2nd (26)	34	42	19	0	04
Metro. Num. 2nd (26)	38	0	30	0	30
Scale Num. 3rd (27)	10	37	26	0	26
Metro. Num. 3rd (27)	26	22	04	44	04
Scale Money 1st (26)	15	19	0	42	23
Metro. Money 1st (26)	30	30	11	0	27
Scale Money 2nd (27)	04	48	10	30	07
Metro. Money 2nd (27)	0	07	0	41	51
Scale Money 3rd (27)	0	10	33	33	22
Metro. Money 3rd (27)	0	0	30	07	63

TABLE 23 (continued)

PERCENTAGES OF PREDICTIONS OF PUPIL POSITION IN THE OAKLEAF CURRICULUM SEQUENCE: SCALED TEST RAW SCORE AND METROPOLITAN TEST RAW SCORE

Test and Grade	Percent Perfect	Percent One Skill Off	Percent Two Skills Off	Percent Three Skill Off	Percent More Than Three Skills Off
Scale Time 1st (25)	04	56	0	16	24
Metro. Time 1st (25)	0	80	0	20	0
Scale Time 2nd (26)	0	38	11	30	19
Metro. Time 2nd (26)	80	04	0	04	11
Scale Time 3rd (27)	0	33	26	26	15
Metro. Time 3rd (27)	26	07	0	15	51

The results appeared inconsistent. The scale tests made better predictions in some instances whereas the Metropolitan tests were better in others. To attempt an explanation for these inconsistencies, the number of different items employed to test the various levels in each unit were determined. As discussed in the section devoted to the validity of the scaled tests, a greater variety of items covering a range of behavioral objectives should enhance the predictions of pupil position in the curriculum sequence. The same was suggested as a reason for the differences between the scaled and Metropolitan tests. As a result it was necessary to obtain the number of items testing each level of the five units. The scaled tests covered all three grades, but each of the Metropolitan batteries only covered one grade. Therefore, in order to compare the two types of tests the range of levels in which the pupils of each grade were working were also obtained. Only the items included within the range of levels for a particular grade were counted. The number of items testing each level of the five units are presented in Table 24. The braces above the levels represent the ranges of level for each grade.

TABLE 24

NUMBER OF ITEMS TESTING EACH LEVEL OF THE FIVE SELECTED UNITS IN THE OAKLEAF CURRICULUM: SCALED TESTS VERSUS METROPOLITAN TESTS

Subtraction

				3rd Grade	
			2nd Grade		
	1st Grade				
	A	B	C	D	E
Scaled Test Items	1	3	3	4	0
Metro. I Items	0	3	0	0	0
Metro. II Items	0	3	3	2	0
Metro. III Items	0	0	4	2	2

Addition

				3rd Grade	
			2nd Grade		
	1st Grade				
	A	B	C	D	E
Scaled Test Items	1	2	5	2	4
Metro. I Items	0	3	0	0	0
Metro. II Items	0	1	5	1	0
Metro. III Items	0	1	5	3	3

Numeration

				3rd Grade	
			2nd Grade		
	1st Grade				
	A	B	C	D	E
Scaled Test Items	5	4	2	3	0
Metro. I Items	3	4	1	0	0
Metro. II Items	2	4	2	0	0
Metro. III Items	0	0	0	0	3

Money

				3rd Grade	
			2nd Grade		
	1st Grade				
	A	B	C	D	E
Scaled Test Items	1	2	2	6	1
Metro. I Items	0	1	2	0	0
Metro. II Items	0	0	1	1	2
Metro. III Items	0	0	0	0	4

TABLE 24 (continued)

NUMBER OF ITEMS TESTING EACH LEVEL OF THE FIVE SELECTED UNITS IN THE OAKLEAF CURRICULUM: SCALED TESTS VERSUS METROPOLITAN TESTS

<table>
<tr><th>Time</th><th></th><th></th><th></th><th colspan="2">3rd Grade</th></tr>
<tr><th></th><th></th><th></th><th colspan="3">2nd Grade</th></tr>
<tr><th></th><th colspan="4">1st Grade</th><th></th></tr>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th></tr>
<tr><td>Scaled Test Items</td><td>0</td><td>3</td><td>9</td><td>4</td><td>0</td></tr>
<tr><td>Metro. I Items</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>Metro. II Items</td><td>0</td><td>0</td><td>3</td><td>0</td><td>0</td></tr>
<tr><td>Metro. III Items</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
</table>

In considering tables 23 and 24 together, only eight of the fifteen comparisons could be explained by tests having a greater variety of items to cover a range of objectives. The remaining seven comparisons could not be explained with the above reasoning. Closer inspection of the data yielded several explanations.

For first grade Subtraction the scaled test had eleven items to the Metropolitan's three, yet the two tests were about equal in predicting the pupils within three skills of their actual unit and skill. The Metropolitan test, however, predicted sixty-three percent of the pupils within one skill while the scaled test only predicted forty percent. There were seven items in the scaled test which were at levels C and D. The last objective tested by the Metropolitan test was the last objective in level B. The majority of the first grade students (nineteen) did not work at all in the Subtraction unit, but were placed at the beginning of level C for 1965. These pupils, however, were able to answer items in the scaled test which pertained to a higher level, even though they had not reached that level in the curriculum sequence. Because of this the pupils were predicted above where they were placed. Such occurred in twelve cases. No opportunity to pass such items was present in the Metropolitan test. Therefore, all those passing the three items were predicted at the first skill of level C.

The same line of reasoning was suggested as an explanation for the results of the second grade Addition tests, the first grade Money tests, and the first and second grade Time tests.

The absence of items at level E for the scaled Numeration test, third grade, was suggested as a reason for this test having a greater number of predictions more then three skills off when compared to the Metropolitan test. Only three of the third grade pupils were working in level D. All others were working above that level. While only three items constituted the Metropolitan Numeration test, all were at level E.

One suggested reason for the difference between the third grade Addition tests was that the items at level E on the scaled test were easier items than those at level E on the Metropolitan. The level E Metropolitan included both the addition of four-digit numerals with carrying and statement problems involving carrying of multiple-digit numerals. Neither of these objectives was tested on the scaled tests.

Some credence was given this suggestion because twenty level E predictions were made from the scaled test and eighteen of these were too high. Only nine level E predictions were made from the Metropolitan with six being too high.

VI. CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

From the results it can be concluded that it is indeed possible to construct sequentially scaled achievement tests in certain areas of arithmetic. No conclusions were reached concerning the Time tests, since further revision was necessary before it would scale. Several other tests, while scaled, require improvement. These results pertain to five selected areas in arithemetic covering grades one through three; thus the conclusions are restricted to these areas. Further investigation at all grade levels in arithmetic achievement areas such as multiplication, division, and fractions is warranted. The nature of arithmetic lends itself to scaling, since many processes depend on previously learned skills. The scaling procedures should also be attempted in other subject matter areas to determine the scope of application.

The application of any particular scaled test may be limited, however. The results of the Numeration test showed that a test may scale for one group and not another. The rank difference correlations between the tests for the Oakleaf and Sickman subjects indicated fluctuation in the item orders between groups. This result meant that a given order of items may scale for one group but need rearrangement to scale for another group. Therefore, when describing a test as scaled the statement should pertain to a specific group and order of items. A reason suggested for the variation in item orders is that two different types of schools were employed in the present investigation and the orders in which the behavioral objectives were taught may have not been the same. This suggested that while the scaling methodology may have a wide range of application, any specific scaled test may have a rather restricted range of application dependent on the order of the objectives in the curriculum. It would be hazardous, therefore, to attempt to infer behavior from total score for a group whose curriculum sequence differed from the sequence on which the scaled test was based.

These results also substantiated the warnings of Schuessler, Torgerson, Campbell, and Kerckhoff (see page 26) who stated that one could not conclude, on the basis of a sample, that a universe was scaled. In the present investigation the same items had a different scaled order for two different samples. This suggests that certain skills in mathematics may not be prerequisite to others but, rather, depend on the order in which they are taught. The scaling procedure, therefore, may have application in the determination of prerequisite skills. Such application should be the topic of future research.

The final orders of items which obtained the maximum criteria for scales in the Oakleaf and Sickman samples, respectively were not the same orders as logically postulated for the objectives. This suggested that empirical verification through the scaling methodology should be attempted before an order of objectives and their corresponding items are considered scaled.

The resultant item orders also indicated that the objectives in the Oakleaf curriculum were not in the best sequence. For example, in both the Oakleaf Subtraction and Addition tests the objectives and items involving borrowing or carrying with single or multiple-digit numerals were more difficult than those which involved multiple-digit numerals without either borrowing or carrying. In each instance, however, borrowing and carrying with single-digit numerals was taught at the level before multiple-digit numerals without borrowing or carrying. Such examples suggested that the use of the scaling methodology may have application in curriculum analysis. Such an application should be the topic of future research.

The scaled tests were constructed as one method of obtaining greater meaning from test raw score, namely, to be able to infer behavior from raw score. The scaled test raw scores should indicate what specific behaviors a pupil has mastered on the test. Excluding the results of the Time test which was not scaled, such inferences were possible from the tests obtained in the present investigation. In each test from eighty-seven to ninety-seven percent of the total raw scores were within one item of representing the first n items passed.

These results concerning the items represented by the total score were in contrast to the results from the total raw scores of a norm-referenced test, the Metropolitan Achievement Test. Except to state that so many items were passed and failed, no meaning could be given to the total raw scores of the two arithmetic subtests of the Metropolitan at each of the first three grades. These total raw scores included items from Addition, Subtraction, Multiplication, Division, Fractions, Money, Time, etc., but there was no way of telling from the total score which of these items were passed and which were failed. Even with the Metropolitan items identified according to units, the scaled tests had more stable score patterns. This latter result should be further substantiated, however, for it was based on only two scaled tests, Addition and Subtraction.

The items which appeared in the scaled tests and in the Metropolitan tests were similar in many cases. When the scaling methodology was applied to the unit tests derived from the Metropolitan two scaled tests were obtained. The apparent reason for more scaled tests not being obtained was lack of item rather than differences in items. The essential point was that scaled tests could be obtained from a norm-referenced test. This suggested that the two types of tests, scaled and norm-referenced, differed mainly because of the nature of the information desired from the raw score.

The results of application of two item analysis procedures to the scaled tests suggested that the upper-lower twenty-seven percent method was inappropriate for scaled tests since items rejected by this procedure were good items for a scale. The results also further substantiated the contention that the type of item analysis applied should depend on the type of test desired. Further research is needed in this area with the investigation covering a wider range of item analysis procedures and varying types of tests.

The split-half reliability coefficients obtained for the scaled tests were of the same magnitude as those of the Metropolitan subtests even though the procedure was more suited to the Metropolitan. Because of the contaminating factors the scaled tests may well have higher split-half reliability than the Metropolitan arithmetic subtests.

Further research is suggested in the area of test-retest reliability for scaled tests. One test-retest procedure which could be attempted would be to administer the scaled test, wait a day, and re-administer the test. This should practically eliminate the contaminating factor of having the ranks of students change. If extended periods of time are allowed to pass between administrations, a procedure might be developed which would encompass only the items passed on the initial test administration. That is, if a pupil passed the first n items the initial time he should have passed at least the first n items the second time.

The validity procedure which involved predictions of pupil position in the curriculum sequence is suggested as providing a type of construct validity. The results suggested that this type of validity can be increased by greater coverage of the behavioral objectives in a curriculum sequence. While this conclusion should have been obvious, the results afforded empirical evidence. The results also suggested certain contaminating factors for this type of validation: (1) If the objectives are not in the same sequence in the curriculum as they are in the test, the resultant predictions may be less accurate than if the orders were the same. (2) Two of the factors appeared to interact: the number of items and the number of objectives tested. It is suggested that this type of validity will be improved with more objectives in a given level being assessed with a sufficient number of items. Further research is needed, however, to determine the amount of interaction of the above factors and to determine how many items constitute a sufficient number to test an objective.

To employ the above validity procedure, however, a daily record of pupil achievement and progress in the curriculum sequence is required. Future research in this area could involve the evaluation of this validity procedure as a type of construct validity, and also the evaluation, employing this validity procedure, of tests currently being employed in the schools as measures of achievement.

In addition to consideration of validity, future investigation should employ caution in the construction of scaled tests. When constructing scaled tests it should be made certain that all groups to be tested follow the same curriculum sequence, or adjust for differences being made by scaling for each group separately.

If behavior is to be inferred from test score both an individual's responses and the order of items should be consistent. Therefore, future investigations should consider both types of reliability, stability of test scores and stability of item orders, when a scaled test is evaluated. Caution should also be exercised in the selection of item analysis procedures since the type of test desired may well dictate the item analysis procedure to be employed.

It is also suggested that the scaling methodology has application in the schools for use by teachers and curriculum designers. The methodology is not complicated, and therefore, should be readily accessible to the classroom teacher. Given the objectives a teacher who can write good test items should find the methodology useful for placement, diagnostic, and achievement testing.

From the results of the present investigation those concerned with curriculum design should also find application for the scaling methodology. The methodology may be employed as a device for analyzing curriculum sequences or for determining certain prerequisite skills.

BIBLIOGRAPHY

Books

Carroll, John B. "Criteria for the Evaluation of Achievement Tests--from the Point of View of Their Internal Statistics," *Proceedings, 1950 Invitational Conference on Testing Problems*, pp. 95-99. Princeton: Educational Testing Service, 1951.

Ebel, Robert L. *Measuring Educational Achievement.* Englewood Cliffs: Prentice-Hall, Inc., 1965.

Flanagan, John C. "Units, Scores, and Norms," *Educational Measurement*. Edited by E. F. Lindquist. Washington, D.C.: American Council on Education, 1951.

Guilford, J. P. *Fundamental Statistics in Psychology and Education.* Third Edition. New York: McGraw-Hill Book Company, 1956.

Guttman, Louis. "The Principal Components of Scalable Attitudes," *Mathematical Thinking in the Social Sciences.* Edited by Paul F. Lazarsfeld. Glencoe, Illinois: Free Press, 1954.

BIBLIOGRAPHY

Guttman, Louis. "The Basis for Scalogram Analysis" and "The Problem of Attitude and Opinion Measurement," in Samuel A. Stouffer *et al.* *Measurement and Prediction.* Vol. IV of *Studies in Social Psychology in World War II.* 4 vols. Princeton: Princeton University Press, 1950.

Lindvall, C. M. *Testing and Evaluation: An Introduction.* New York: Harcourt, Brace and World, Inc., 1961.

Suchman, Edward A. "The Scalogram Board Technique for Scale Analysis," in Samuel A. Stouffer *et al.* *Measurement and Prediction.* Vol. IV of *Studies in Social Psychology in World War II.* 4 vols. Princeton: Princeton University Press, 1950.

Torgerson, Warren S. *Theory and Methods of Scaling.* New York: John Wiley & Sons, Inc., 1958.

Tucker, Ledyard R. "Selecting Appropriate Score Scales for Tests--Scales Minimizing the Importance of Reference Groups," *Proceedings, 1952 Invitational Conference on Testing Problems.* Princeton: Educational Testing Service, 1953.

BIBLIOGRAPHY

Books

Carroll, John B. "Criteria for the Evaluation of Achievement Tests--from the Point of View of Their Internal Statistics," *Proceedings, 1950 Invitational Conference on Testing Problems*, pp. 95-99. Princeton: Educational Testing Service, 1951.

Ebel, Robert L. *Measuring Educational Achievement*. Englewood Cliffs: Prentice-Hall, Inc., 1965.

Flanagan, John C. "Units, Scores, and Norms," *Educational Measurement*. Edited by E. F. Lindquist. Washington, D.C.: American Council on Education, 1951.

Guilford, J. P. *Fundamental Statistics in Psychology and Education*. Third Edition. New York: McGraw-Hill Book Company, 1956.

Guttman, Louis. "The Principal Components of Scalable Attitudes," *Mathematical Thinking in the Social Sciences*. Edited by Paul F. Lazarsfeld. Glencoe, Illinois: Free Press, 1954.

Guttman, Louis. "The Basis for Scalogram Analysis" and "The Problem of Attitude and Opinion Measurement," in Samuel A. Stouffer *et al*. *Measurement and Prediction*. Vol. IV of *Studies in Social Psychology in World War II*. 4 vols. Princeton: Princeton University Press, 1950.

Lindvall, C. M. *Testing and Evaluation: An Introduction*. New York: Harcourt, Brace and World, Inc., 1961.

Suchman, Edward A. "The Scalogram Board Technique for Scale Analysis," in Samuel A. Stouffer *et al*. *Measurement and Prediction*. Vol. IV of *Studies in Social Psychology in World War II*. 4 vols. Princeton; Princeton University Press, 1950.

Torgerson, Warren S. *Theory and Methods of Scaling*. New York: John Wiley & Sons, Inc., 1958.

Tucker, Ledyard R. "Selecting Appropriate Score Scales for Tests--Scales Minimizing the Importance of Reference Groups," *Proceedings, 1952 Invitational Conference on Testing Problems*. Princeton: Educational Testing Service, 1953.

Tyler, Fred T. and Walter R. Stellwagen. "The Search for Evidence about Individual Differences," Individualizing Instruction. Sixty-first Yearbook of the National Society for the Study of Education, Part I. Chicago: The University of Chicago Press, 1962.

Tyler, Ralph W. "The Development of Instruments for Assessing Educational Progress," Proceedings of the 1965 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1966.

Periodicals

Abell, Helen C. "The Use of Scale Analysis in a Study of Differential Adoption of Homemaking Practices," Rural Sociology, XVII (1952), pp. 161-167.

Auld, Frank Jr., L. A. Eron, and J. Laffal. "Application of Guttman's Scaling Method to the T.A.T.," Educational and Psychological Measurement, XV (1955), pp. 422-435.

Bligh, Harold F. "Empirical Investigation of Methods of Scaling Achievement Tests Based on Interelationship of Items," Dissertation Abstracts, XIX (1958), pp. 2648-2649.

Borgatta, Edgar F. and David G. Hays. "Some Limitations on the Arbitrary Classification of Non-Scale Response Patterns in a Guttman Scale," Public Opinion Quarterly, XVI (1952), pp. 410-416.

Brown, C. W., P. Bartelme, and G. M. Cox. "The Scoring of Individual Performance on Tests Scaled According to the Theory of Absolute Scaling," Journal of Educational Psychology, XXIV (1933), pp. 654-662.

Campbell, Ernest Q. and Alan C. Kerckhoff. "A Critique of the Concept 'Universe of Attributes,'" Public Opinion Quarterly, XXI (1957), pp. 295-303.

Clark, K. E. and P. H. Kriedt. "An Application of Guttman's New Scaling Techniques to an Attitude Questionnaire," Educational and Psychological Measurement, VIII (1948), pp. 215-223.

Coughenour, C. M. and J.R. Christiansen. "Farmers' Knowledge: An Appraisal of Stouffer's H-technique," Rural Sociology, XXIII (1958), pp. 253-262.

Coulson, John E. and John F. Cogswell. "Effects of Individualized Instruction on Testing," Journal of Educational Measurement, II (1965), pp. 59-64.

Davis, James A. "On Criteria for Scale Relationships," *American Journal of Sociology*, LXIII (1958), pp. 371-380.

Dodd, S. C. " A Simple Test for Predicting Opinions from Their Subclasses," *International Journal of Opinion and Attitude Research*, II (1948), pp. 1-25.

Ebel, Robert L. "Content Standard Test Scores, "*Educational and Psychological Measurement*, XXII (1962), pp. 15-25.

Edwards, Allen L. "On Guttman's Scale Analysis," *Educational and Psychological Measurement*, VII (1948), pp. 313-318.

Edwards, Allen L. and Franklin P. Kilpatrick. "A Technique for the Construction of Attitude Scales," *Journal of Applied Psychology*, XXXII (1948), pp. 374-384.

Edwards, Allen L. and Franklin P. Kilpatrick. "Scale Analysis and the Measurement of Social Attitudes," *Psychometrika*, XIII (1948), pp. 99-114.

Eysenck, H. J. "Measurement and Prediction; A Discussion of Volume IV of Studies in Social Psychology in World War II: I.," *International Journal of Opinion and Attitude Research*, V (1951), pp. 95-102.

Eysenck, H. J. and S. Crown. "An Experimental Study in Opinion-Attitude Methodology, "*International Journal of Opinion and Attitude Research*, III (1949), pp. 47-86.

Festinger, Leon. "The Treatment of Qualitative Data by 'Scale Analysis,'" *Psychological Bulletin*, XLIV (1947), pp. 149-161.

Gibson, Wilfred A. "A Simple Procedure for Rearranging Matrices," *Psychometrika*, XVIII (1953), pp. 111-113.

Glaser, Robert. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," *American Psychologist*, XVIII (1963), pp. 519-521.

Glaser, Robert. "Multiple Operation Measurement," *Psychological Review*, LVII (1950), pp. 241-253.

Glaser, Robert. "The Application of the Concepts of Multiple-Operation Measurement to the Response Patterns on Psychological Tests," *Educational and Psychological Measurement*, XI (1951), pp. 372-382.

Grossnickle, Louise R. "The Scaling of Test Scores by the Method of Paired Comparisons," Psychometrika, VII (1942), pp. 43-64.

Guttman, Louis. "A Basis for Scaling Qualitative Data," American Sociological Review, IX (1944), pp. 139-150.

Guttman, Louis. "Measurement and Prediction; a Discussion of Vol. IV of Studies in Social Psychology in World War II: II. Scale Analysis, Factor Analysis, and Dr. Eysenck," International Journal of Opinion and Attitude Research, V (1951), pp. 103-120.

Guttman, Louis. "On Festinger's Evaluation of Scale Analysis," Psychological Bulletin, XLIV (1947), pp. 451-465.

Guttman, Louis. "On Smith's Paper on 'Randomness of Error' in Reproducible Scales," Educational and Psychological Measurement XIII (1953), pp. 505-511.

Guttman, Louis. "The Cornell Technique for Scale and Intensity Analysis," Educational and Psychological Measurement, VII (1947), pp. 247-280.

Kofsky, Ellin. "Developmental Scalogram Analysis of Classificatory Behavior," Dissertation Abstracts, XXIV (1963), p. 2576.

Kriedt, P. H. and K. E. Clark. "'Item Analysis' Versus 'Scale Analysis,'" Journal of Applied Psychology, XXXIII (1949), pp. 114-121.

Lesser, Gerald A. "Application of Guttman's Scaling Method to Aggressive Fantasy in Children," Educational and Psychological Measurement, XVIII (1958), pp. 543-551.

Loevinger, Jane. "A Systematic Approach to the Construction and Evaluation of Tests of Ability," Psychological Monographs, LXI (1947), No. 285.

Loevinger, Jane. "The Technic of Homogeneous Tests Compared with some Aspects of 'Scale Analysis' and Factor Analysis," Psychological Bulletin, XLV (1948), pp. 507-529.

Lord, Frederic. "Scaling," Review of Educational Research, XXIV (1954), pp. 375-392.

McGinnis, Robert, "Scaling Interview Data," American Sociological Review, XVIII (1953), pp. 514-521.

Menzel, Herbert. "A New Coefficient for Scalogram Analysis," *Public Opinion Quarterly*, XVII (1953), p. 268-280.

Niven, Jarold R. "A Comparison of Two Attitude Scaling Techniques," *Educational and Psychological Measurement*, XIII (1953), pp. 65-76.

Pearson, Richard G. "Plus Percentage Ratio and the Coefficient of Scalability," *Public Opinion Quarterly*, XXI (1957), pp. 279-380.

Postove, Mary Jane. "Selection of Items for a Speech-Reading Test by Means of Scalogram Analysis," *Journal of Speech and Hearing Disorders,* XXVII (1962), pp. 71-75.

Schuessler, Karl. "Item Selection in Scale Analysis," *American Sociological Review*, XVII (1952), pp. 183-192.

Scott, John Finley. "Two Dimensions of Deliquent Behavior," *American Sociological Review*, XXIV (1959), pp. 240-243.

Siegel, Arthur I. and Douglas G. Schultz, "Thurstone and Guttman Scaling of Job Related Technical Skills, *Psychological Reports*, X (1962), pp. 855-866.

Smith, R. G., Jr. "'Randomness of Error' in Reproducible Scales," *Educational and Psychological Measurement*, XI (1951), pp. 587-596.

Stouffer, Samuel A. *et al*. "A Technique for Improving Cumulative Scales," *Public Opinion Quarterly*, XVI (1952), pp. 273-291.

Suchman, Edward A. "The Logic of Scale Construction," *Educational and Psychological Measurement*, X (1950), pp. 79-93.

Tucker, Ledyard R. "A Level of Proficiency Scale for a Unidimensional Skill," *American Psychologist*, VII (1952), 408. (Abstract)

White, Benjamin W. and Eli Saltz. "Measurement of Reproducibility," *Psychological Bulletin*, LIV (1957), pp. 81-99.

Unpublished Materials

Cox, Richard C. and Glenn T. Graham. "The Development of a Sequentially Scaled Achievement Test." Paper read at the 50th Annual Meeting of the American Educational

Research Association, Chicago, Illinois, February 17, 1966.

Cox, Richard C. and Julie S. Vargas. "A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests." Paper read at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, February 17, 1966.

Graham, Glenn T. and Richard C. Cox. "An Attempt to Determine the Scalability of an Elementary Math Achievement Test." Paper read at the Pennsylvania Educational Research Association Conference, Pittsburgh, Pennsylvania, April, 1965.

Husek, T. R. "Different Kinds of Evaluation and their Implications for Test Development." Paper read at the 50th Annual Meeting of the American Educational Research Association, Chicago, Illinois, February 19, 1966.

APPENDICES

APPENDIX A

OBJECTIVES AND SAMPLE TEST ITEMS FOR PILOT STUDY

Objectives	Sample Test Items
The student is able to:	
1. Recognize numerals from 1 to 10.	1. 1 2 3 4 "Draw a circle around the 2."
2. a) Determine which numeral comes before or after another numeral.	2. a) 10 8 5 2 "Draw a circle around the number that comes just after 7."
b) Determine which of 2 numerals is the largest or smallest.	b) 7 5 "Draw a circle around the largest numeral."
3. Discriminate between +, -, =, ≠.	3. + - = ≠ "Draw a circle around the sign which means to add."
4. a) Add two single-digit numerals with sums to 10, vertically.	4. a) $\begin{array}{r} 5 \\ +4 \\ \hline \end{array}$
b) Add two single-digit numerals with sums to 10, horizontally.	b) 3 + 1 =
5. a) Add two single-digit numerals involving carrying, horizontally.	5. a) 8 + 3 =
b) Add two single-digit numerals involving carrying, vertically.	b) $\begin{array}{r} 7 \\ +8 \\ \hline \end{array}$

Objectives	Sample Test Items
6. a) Add three single-digit numerals involving carrying, vertically.	6. a) $\begin{array}{r} 1 \\ 8 \\ +4 \\ \hline \end{array}$
b) Add three single-digit numerals involving carrying, horizontally.	b) 7 + 3 + 2 =
7. a) Place one and two-digit numerals in a column so they could be added.	7. a) 15 16 2 "Place these numerals in a column so they could be added."
b) Determine which column of numerals is written so they could be added.	b) $\begin{array}{r} 15 \\ 16 \\ \underline{2} \end{array} \quad \begin{array}{l} 15 \\ 16 \\ \underline{2} \end{array} \quad \begin{array}{r} 15 \\ 12 \\ \underline{2} \end{array}$ "Draw a circle around the column which is written so it could be added.
8. Add 2 two-digit numerals without carrying.	8. $\begin{array}{r} 20 \\ +11 \\ \hline \end{array}$
9. Add 2 three-digit numerals without carrying.	9. $\begin{array}{r} 215 \\ +723 \\ \hline \end{array}$
10. Add 2 two-digit numerals with carrying.	10. $\begin{array}{r} 58 \\ +36 \\ \hline \end{array}$

APPENDIX B

MONEY OBJECTIVES

The student will be able to:

1. Identify pennies, nickels, and dimes.
2. Give the value of a penny, nickel, or dime in cents.
3. Identify the dollar and cent signs.
4. Give the value of a combination of two coins (penny, nickel, and dime).
5. Give the value of a combination of 3 or more coins (penny, nickel, dime).
6. Identify quarters, half-dollars, and dollars.
7. Identify equivalent amounts of money.
8. Give the number of pennies, nickels, dimes, quarters, and half-dollars in a dollar.
9. Give the value of a sum of money expressed in decimal notation.
10. Add two amounts of money expressed in decimal notation, without carrying.
11. Add two amounts of money expressed in decimal notation, with carrying.
12. Can make change for amounts of money of $5.00 or less.

ADDITION OBJECTIVES

The student will be able to:

1. Count the number of objects in a set (less than 10).
2. Add two single-digit numerals with sums less than 10.
    a. horizontally arranged

  - b. vertically arranged

    (Where not specified the numerals are arranged vertically)

3. Add two single-digit numerals with sums greater than or equal to 10.
   - a. horizontally arranged
   - b. vertically arranged
4. Add three single-digit numerals.
5. Add two two-digit numerals without carrying.
6. Add three two-digit numerals without carrying.
7. Identify the proper way to arrange numerals so that they could be added.
8. Add two three-digit numerals without carrying.
9. Add three three-digit numerals without carrying.
10. Add two two-digit numerals with carrying.
11. Add two three-digit numerals with carrying.
12. Add three two-digit numerals with carrying.

SUBTRACTION OBJECTIVES

The student will be able to:

1. Subtract one group of objects from another, 10 objects or less.
2. Subtract using the expression "take away," with numerals less than 10.
3. Subtract single-digit numerals.
   - a. horizontally arranged
   - b. vertically arranged

     (Where not specified the numerals are arranged vertically)

4. Subtract a single-digit numeral from a two-digit numeral (less than 20) without borrowing.
5. Identify the proper way to arrange numerals so that they can be subtracted.
6. Subtract a single-digit numeral from a two-digit numeral (less than 20) with borrowing.
7. Subtract two two-digit numerals without borrowing.
8. Subtract a two-digit numeral from a three-digit numeral without borrowing.
9. Subtract two three-digit numerals without borrowing.
10. Subtract two two-digit numerals with borrowing.

TIME TELLING OBJECTIVES

The student will be able to:

1. Fill in the missing numerals on a clock face.
2. Identify the hour and minute hands.
3. Tell time to the hour.
4. Tell time to the half-hour.
5. Tell time to the quarter-hour.
6. Count the number of minutes between two points on a clock.
    a. up to 30 minutes
    b. from 30 minutes to 1 hour
7. Discriminate between clock faces when the time is expressed as:
    a. _____ minutes after _____ (half-hour)
    b. _____ thirty (half-hour)

8. Write the time, in numerals (e.g. 8:00), when given a time.

9. Write the time in numerals when given a clock face:

   a. to the hour
   b. to the half-hour
   c. to the quarter-hour
   d. to five-minute intervals
   e. to the minute

NUMERATION OBJECTIVES

The student will be able to:

1. Recognize numerals from 1 to 10.
2. Write numerals sequentially from 1 - 10.
3. Find the larger or smaller of two numerals from 1 - 10.
4. Write the numeral that comes just after a numeral from 1 - 10.
5. Write the numeral that comes just before a numeral from 1 - 10.
6. Recognize numerals from 1 - 100.
7. Write numerals sequentially that come between two given numerals from 1 - 100.
8. Find the larger or smaller of two numerals from 1 - 100.
9. Count the number of objects in a group presented visually.
10. Count by fives from 1 - 100.
11. Write the numerals that come before and after a given number or series of numbers from 1 - 100.
12. Write sequentially, counting by ones, the numerals that follow a given numeral from 100 - 1000.
13. Write the numeral that comes just after a given numeral from 100 - 1000.
14. Write the numeral that comes just before a given numeral from 100 - 1000.

APPENDIX C

A SHORT DESCRIPTION OF MATHEMATICS UNITS

1. A Numeration - Counting to ten.
2. A Addition - Addition to sums of six with pictured objects.
3. A Fractions - Identification of 1/2 of objects and small sets.
4. A Money - Recognition of common coins (penny, nickel, dime).
5. A Time - The day as a unit of time.
6. A Systems of Measurement - Qualitative dimensional discrimination by verbal directions.
7. A Geometry - Recognition of simple geometric figures.
8. B Numeration - Counting to 100. Use of ordinals to 10th.
9. B Addition - Addition to sums of 10.
10. B Money - Beginning money equivalents (5¢ = 1 nickel).
11. B Time - Clock reading to the hour.
12. B Systems of Measurement - Beginning equivalent length (3 ft. = 1 yd.).
13. B Geometry - Draws simple geometric figures.
14. C Numeration - Counting to 150.
15. C Place Value - Place value charting to hundreds.
16. C Addition - Two digit sums without carrying but with expanded notation.
17. C Subtraction - Two digit differences without carrying but with expanded notation.
18. C Combination of Processes - Word problems with skills learned to this point plus selection of proper operation to solve problems.
19. C Fractions - With fractions to 1/4 divides single objects and groups of objects.

20. C Money - Practical use of penny, nickel, dime, and quarter.

21. C Time - Solves problems requiring addition or subtraction of hours.

22. C Systems of Measurement - Converts units: inches - feet, pint - quart - cup, dozen - 1/2 dozen.

23. C Geometry - Recognizes and names solid geometric figures.

24. C Special Topics - Reads Roman numerals, to 10; reads thermometer; reads charts and graphs.

25. D Numeration - Counting to 1,000 (reading and writing numerals with skip counting.).

26. D Place Value - Makes and reads place value charts to thousands.

27. D Addition - Begins addition with carrying.

28. D Subtraction - Begins subtraction with borrowing.

29. D Multiplication - Does multiplication as repeated addition. Memorizes tables through 5 x 5.

30. D Division - Does division as partition, inverse to addition, and memorizes tables through 25 divided by 5.

31. D Combination of Processes - Solves problems requiring selection and discrimination of many processes.

32. D Fractions - Applies fractional concepts (2/3, 3/4) to objects and groups. Begins formal operations (1/2 x 8 = ?).

33. D Money - Operates with money values to $5.00.

34. D Time - Tells time to the minute and uses time in problems.

35. D Systems of Measurement - Extends linear and volume systems and begins metric system with centimeters.

36. D Geometry - Identified open versus closed curves, line segments versus lines.

37. D Special Topics - Reads Roman numerals to 30.

38. E Numeration - Identifies odd versus even numbers; rounds and estimates numbers.

39. E Place Value - Uses place value to millions; begins exponents of base 10.

40. E Addition - Performs addition with carrying to thousands.

41. E Subtraction - Does subtraction with borrowing to hundreds.

42. E Multiplication - Does multiplication as repeated addition. Uses associative and distributive principle and does simple multiplication with carrying.

43. E Division - Uses ladder algorithm for division.

44. E Combination of Processes - Solves using n as variable. Does operations with competing processes.

45. E Fractions - Identifies equivalent fractions; adds fraction with a common denominator.

46. E Money - Adds and subtracts money values using decimal notation.

47. E Time - Uses seconds in time problems.

48. E Systems of Measurement - Adds and subtracts measures by regrouping when necessary.

49. E Geometry - Identifies simple line figures (equilateral triangle, quadrilateral, parallel lines, midpoint, end points, right angle, intersecting lines, perpendicular lines).

50. E Special Topics - Uses simple maps.

APPENDIX D

SCALOGRAM OF OAKLEAF SUBTRACTION TEST

	1	2	3	4	5	6	7	8	9	10	11
1	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	
3	X	X	X	X	X	X	X	X	X	X	
4	X	X	X	X	X	X	X	X	X	X	
5	X	X	X	X	X	X	X	X	X	X	
6	X	X	X	X	X	X	X	X	X	X	
7	X	X	X	X	X	X	X	X	X	X	
8	X	X	X	X	X	X	X	X	X	X	
9	X	X	X	X	X	X	X	X	X	X	
10	X	X	X	X	X	X	X	X	X	X	
11	X	X	X	X	X	X	X	X	X	X	
12	X	X	X	X	X	X	X	X	X	X	
13	X	X	X	X	X	X	X	X	X	X	
14	X	X	X	X	X	X	X	X	X	X	
15	X	X	X	X	X	X	X	X	X	X	
16	X	X	X	X	X	X	X	X	X	X	
17	X	X	X	X	X	X	X	X		X	X
18	X	X	X	X	X	X	X	X	X		
19	X	X	X	X	X	X	X	X	X		
20	X	X	X	X	X	X	X	X	X		
21	X	X	X	X	X	X	X	X	X		
22	X	X	X	X	X	X	X	X	X		
23	X	X	X	X	X	X	X	X	X		
24	X	X	X	X	X	X	X	X	X		
25	X	X	X	X	X	X	X	X	X		
26	X	X	X	X	X	X	X	X	X		
27	X	X	X	X	X	X	X	X	X		
28	X	X	X	X	X	X	X	X	X		
29	X	X	X	X	X	X	X	X	X		
30	X	X	X	X	X	X	X	X	X		
31	X	X	X	X	X	X	X	X		X	
32	X	X	X	X	X	X	X	X		X	
33	X	X	X	X	X	X	X	X		X	
34	X	X	X	X	X	X	X	X		X	
35	X	X	X	X	X		X	X	X	X	
36	X	X	X	X	X	X	X	X			X
37	X	X	X	X	X	X	X	X			
38	X	X	X	X	X	X	X	X			
39	X	X	X	X	X		X	X	X		
40	X	X	X	X	X	X	X	X			
41	X	X	X	X	X	X	X	X			
42	X	X	X	X	X	X	X	X			
43	X	X	X	X	X	X	X	X			
44	X	X	X	X	X	X	X	X			
45	X	X	X	X	X	X	X	X			
46	X	X	X	X	X		X	X	X		
47	X	X	X	X	X		X	X		X	
48	X	X		X	X	X	X	X	X		
49	X	X	X	X		X	X	X			

SCALOGRAM OF OAKLEAF SUBTRACTION TEST

	1	2	3	4	5	6	7	8	9	10	11
50	X	X	X	X	X		X	X			
51	X	X	X	X	X	X	X				
52	X	X	X	X	X	X	X				
53	X	X	X		X	X	X	X			
54	X	X	X		X	X	X	X			
55	X	X	X	X	X	X					
56	X	X	X	X	X	X					
57	X	X	X	X	X	X					
58	X	X	X	X		X		X			
59	X	X	X			X	X	X			
60	X	X		X	X	X	X				
61	X	X	X	X	X	X					
62	X	X	X		X		X		X		
63		X	X	X	X	X					
64	X	X	X	X	X						
65	X	X	X	X	X						
66	X	X	X	X			X				
67	X	X	X		X	X					
68	X	X	X			X					
69	X	X		X		X					
70	X	X		X		X					
71	X	X		X		X					
72	X	X	X								
73	X	X		X							
74	X		X								
75				X							

SCALOGRAM OF OAKLEAF SUBTRACTION TEST

	1	2	3	4	5	6	7	8	9	10	11
50	X	X	X	X	X		X	X			
51	X	X	X	X	X	X	X				
52	X	X	X	X	X	X	X				
53	X	X	X		X	X	X	X			
54	X	X	X		X	X	X	X			
55	X	X	X	X	X	X					
56	X	X	X	X	X	X					
57	X	X	X	X	X	X					
58	X	X	X	X		X		X			
59	X	X	X			X	X	X			
60	X	X		X	X	X	X				
61	X	X	X	X	X	X					
62	X	X	X		X		X		X		
63		X	X	X	X	X					
64	X	X	X	X	X						
65	X	X	X	X	X						
66	X	X	X	X			X				
67	X	X	X		X	X					
68	X	X	X			X					
69	X	X		X		X					
70	X	X		X		X					
71	X	X		X		X					
72	X	X	X								
73	X	X		X							
74	X		X								
75				X							